UNIVERSITETI I EVROPËS JUGLINDORE
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА
SOUTH EAST EUROPEAN UNIVERSITY

FAKULTETI I SHKENCAVE DHE TEKNOLOGJIVE BASHKËKOHORE
ФАКУЛТЕТ ЗА СОВРЕМЕНИ НАУКИ И ТЕХНОЛОГИИ
FACULTY OF CONTEMPORARY SCIENCES AND TECHNOLOGIES

THIRD CYCLE OF ACADEMIC STUDIES – DOCTORAL STUDIES

DOCTORAL DISSERTATION TOPIC:

# ASSESSMENT OF THE IMPACT OF DATA QUALITY FOR IMPROVEMENT OF E-SERVICES IN GOVERNMENT INSTITUTIONS

Candidate:

M.Sc. Genc Hamzaj

Mentor:

Prof. Dr. Zamir Dika

Tetovo, September 2022

## Abstract

Provision of data in appropriate level in aspect of quality is one of the key goals for raising the quality of services that will be provided for citizens. Most public administration services, which previously were only able to be realized physically, now can be achieved online since practically every citizen now uses Internet services or has access to the internet, which has huge importance as a necessary condition before government institutions attempt to offer electronic services for citizens.

Due to the increasing number of databases created with the aim to provide electronic services for public administration and the lack of data harmonization or interoperability between these databases, this process caused that the quality of the data has decreased as a result of numerous mistakes that were done and also inconsistencies in the data in these databases.

In addition, due to the extremely high volume of data as well as the numerous diverse data sources and data structures that we have now because of the rapid expansion of IoT devices, evaluating and enhancing the quality of data is becoming very challenging.

The whole process for Assessment and Improvement of Data Quality Approach by the government institutions will be constructed by evaluating the most significant dimensions, metrics, and current most important frameworks, including the relevant assessment and improvement procedures, which is the main objective of this dissertation.

Increasing data quality using different dimensions and adequate approaches is a prerequisite for achieving high quality e-services.

The main dimensions of the qualitative research method that will be implemented in this dissertation with aim to treat the quality of the data are *completeness, uniqueness, timeliness, validity, accuracy, and consistency.*

Due to the extremely huge volume of data and several data sources with various data structures, evaluating and enhancing the quality of data by matching and linking records becomes very challenging. Our focus will be on algorithms that handle large amounts of

data, such as Damerau-Levenshtein distance (DL) algorithm and Levenshtein distance (LV) algorithm.

In order to compare the effectiveness and quality of the data using the specified algorithms, through this dissertation we will conduct experiments in huge datasets with more than 1 million records.

Additionally, we will perform a data cleansing process by analyzing and identifying inaccurate data in datasets, including: order dependency violations, delayed reported issues, anomalous data within certain periods, etc. We will utilize the Power BI tool to evaluate datasets from various sources, making improvements by implementing suitable dimensions and techniques for increasing the quality of the data.

Finally, through this dissertation we will evaluate the e-Services delivered by the Kosovo government portal by using the provided data quality dimensions to evaluate the data quality in the chosen datasets. Additionally, the effectiveness of implementing user-friendly, data quality criteria and dimensions into a single government portal is evaluated with the goal of providing better and improved G2C (Government to Citizens) services. In addition, we will demonstrate a microservice architectural integration model for implementing e-services in the Kosovo government portal.

## Abstrakt

Sigurimi i të dhënave në nivelin e duhur në aspektin e cilësisë është një nga qëllimet kryesore për ngritjen e cilësisë së shërbimeve që do të ofrohen për qytetarët. Pjesa më e madhe e shërbimeve të administratës publike, të cilat më parë mund të realizoheshin vetëm fizikisht, tani mund të arrihen online pasi praktikisht çdo qytetar tani përdor shërbimet e internetit ose ka akses në internet, gjë që ka një rëndësi të madhe si kusht i domosdoshëm përpara se institucionet qeveritare të tentojnë të ofrojnë shërbime elektronike për qytetarët.

Për shkak të rritjes së numrit të bazave të të dhënave të krijuara me qëllim të ofrimit të shërbimeve elektronike për administratën publike dhe mungesës së harmonizimit apo ndërveprimit të të dhënave ndërmjet këtyre bazave të të dhënave. Ky proces bëri që cilësia e të dhënave të ulet si rezultat i gabimeve të shumta që janë bërë si dhe mospërputhjet në të dhënat në këto baza të të dhënave.

Për më tepër, për shkak të vëllimit jashtëzakonisht të lartë të të dhënave, si dhe burimeve të shumta të larmishme të të dhënave dhe strukturave të të dhënave që kemi tani si rezultat i zgjerimit të shpejtë të pajisjeve IoT, vlerësimi dhe rritja e cilësisë së të dhënave po bëhet shumë sfiduese.

I gjithë procesi për Vlerësimin dhe Përmirësimin e Përqasjes së Cilësisë së të Dhënave nga institucionet qeveritare do të ndërtohet duke vlerësuar dimensionet më të rëndësishme, metrikat dhe kornizat më të rëndësishme aktuale, duke përfshirë edhe procedurat përkatëse të vlerësimit dhe përmirësimit, që është objektivi kryesor i këtij disertacioni.

Rritja e cilësisë së të dhënave duke përdorur dimensione të ndryshme dhe qasje adekuate është një parakusht për arritjen e shërbimeve elektronike me cilësi të lartë.

Dimensionet kryesore të metodës së kërkimit cilësor që do të zbatohet në këtë disertacion me qëllim trajtimin e cilësisë së të dhënave janë: *Completeness, Uniqueness, Timeliness, Validity, Accuracy, dhe Consistency.*

Për shkak të vëllimit jashtëzakonisht të madh të të dhënave dhe burimeve të shumta të të dhënave me struktura të ndryshme të dhënash, vlerësimi dhe rritja e cilësisë së të dhënave përmes përputhjes dhe lidhjes së të dhënave bëhet shumë sfiduese. Fokusi ynë do të jetë në algoritmet që trajtojnë sasi të mëdha të dhënash, të tilla si algoritmi i distancës Damerau-Levenshtein (DL) dhe algoritmi i distancës Levenshtein (LV).

Për të krahasuar efektivitetin dhe cilësinë e të dhënave duke përdorur algoritmet e specifikuara, përmes këtij disertacioni do të kryejmë eksperimente në grupe të mëdha të dhënash me më shumë se 1 milion regjistrime.

Për më tepër, ne do të kryejmë procesin e pastrimit të të dhënave duke analizuar dhe identifikuar të dhëna të pasakta në grupet e të dhënave, duke përfshirë shkelja e varësisë së renditjes, problemi i raportimit të vonuar, të dhëna anormale në periudha të caktuara kohore etj. Ne do të përdorim mjetin *Power BI* për të vlerësuar grupet e të dhënave nga burime të ndryshme, duke bërë përmirësime duke zbatuar dimensione dhe teknika të përshtatshme për rritjen e cilësisë së të dhënave.

Së fundi, përmes këtij disertacioni ne do të vlerësojmë shërbimet elektronike të ofruara nga portali i qeverisë së Kosovës duke përdorur dimensionet e ofruara të cilësisë së të dhënave për të vlerësuar cilësinë e të dhënave në grupet e të dhënave të zgjedhura. Për më tepër, efektiviteti i zbatimit të kritereve dhe dimensioneve të cilësisë së të dhënave miqësore për përdoruesit në një portal të vetëm qeveritar vlerësohet me synimin për të ofruar shërbime më të mira dhe të përmirësuara G2C (Qeveria për qytetarët). Përveç kësaj, ne do të demonstrojmë një model të integrimit arkitektonik të mikroshërbimeve për implementimin e shërbimeve elektronike në portalin e qeverisë së Kosovës.

## Апстракт

Обезбедувањето на податоци на соодветно ниво од аспект на квалитет е една од клучните цели за подигнување на квалитетот на услугите што ќе им се даваат на граѓаните. Најголемиот дел од услугите на јавната администрација, кои досега можеа да се реализираат само физички, сега можат да се постигнат преку Интернет, бидејќи практично секој граѓанин сега користи интернет услуги или има пристап до интернет, што има огромна важност како неопходен услов пред владините институции да се обидат да понудат електронски услуги за граѓаните.

Поради зголемениот број на бази на податоци создадени со цел да се обезбедат електронски услуги за јавната администрација и недостатокот на усогласеност на податоците или интероперабилност помеѓу овие бази на податоци, овој процес предизвика намалување на квалитетот на податоците како резултат на многубројните грешки што беа направени а исто така и недоследности во податоците во овие бази на податоци.

Дополнително, поради екстремно високиот обем на податоци, како и бројните разновидни извори на податоци и структури на податоци што ги имаме сега како резултат на брзата експанзија на IoT уредите, оценувањето и подобрувањето на квалитетот на податоците станува многу предизвик.

Целиот процес за проценка и подобрување на пристапот на квалитетот на податоците од страна на владините институции ќе биде конструиран преку евалуација на најзначајните димензии, метрика и актуелните најважни рамки, вклучувајќи ги и соодветните процедури за проценка и подобрување, што е главната цел на оваа дисертација.

Зголемувањето на квалитетот на податоците со користење на различни димензии и соодветни пристапи е предуслов за постигнување висококвалитетни е-услуги.

Главните димензии на методот на квалитативно истражување што ќе се имплементира во оваа дисертација со цел да се третира квалитетот на податоците се *комплетноста, единственоста, навременоста, валидноста, точноста и конзистентноста.*

Поради екстремно огромниот обем на податоци и неколку извори на податоци со различни структури на податоци, оценувањето и подобрувањето на квалитетот на податоците преку усогласување и поврзување на записите станува многу предизвик. Нашиот фокус ќе биде на алгоритми кои ракуваат со големи количини на податоци, како што се алгоритам за растојание Дамерау-Левенштајн (DL) и алгоритам за растојание Левенштајн (LV).

Со цел да се споредат ефективноста и квалитетот на податоците користејќи ги наведените алгоритми, преку оваа дисертација ќе спроведеме експерименти во огромни збирки на податоци со повеќе од 1 милион записи.

Дополнително, ќе извршиме процес на чистење на податоците со анализа и идентификување неточни податоци во збирките на податоци, вклучувајќи: прекршувања на зависноста од нарачки, одложени пријавени проблеми, аномални податоци во одредени периоди итн. Ќе ја користиме алатката *Power BI* за проценка на збирки податоци од различни извори подобрувања со имплементација на соодветни димензии и техники за зголемување на квалитетот на податоците.

Конечно, преку оваа дисертација ќе ги евалуираме е-услугите испорачани од порталот на косовската влада со користење на дадените димензии на квалитетот на податоците за да се оцени квалитетот на податоците во избраните збирки на податоци. Дополнително, се оценува ефективноста на имплементирањето на критериумите и димензиите за квалитет на податоци кои се прифатливи за корисниците и димензии во единствен владин портал со цел да се обезбедат подобри и подобрени G2C (Влада за граѓаните) услуги. Дополнително, ќе демонстрираме модел на архитектонска интеграција на микросервис за имплементација на е-услуги во порталот на косовската влада.

# Declaration

I hereby declare that my Dissertation

**Assessment of the Impact of Data Quality for Improvement of e-Services in Government Institutions**

has been written entirely by myself. This work has not previously been submitted for a degree or diploma in any university.
The research was carried out at the SEEU under the supervision of Prof. Dr. Zamir Dika.

Genc Hamzaj

_____

## Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Dr. Zamir Dika.

His continuous support, advice, and encouragement have made it possible to finish this thesis.

I am especially grateful for his patience in discussing challenges in the context of this thesis and for his advice on how to look at problems from different dimensions.

I would also like to thank my wife Fjolla for her endless patience and love, and my children Rroni and Sparta for their unconditional love and happiness.

Last but not least, I would like to thank my friend Ardian Hoti for support, trust, and encouragement.

# Contents

## List of Figures

## List of Tables

## List of Abbreviations

| Abbreviations | Explanation |
|---|---|
| DQ | Data Quality |
| DQS | Data Quality Services |
| KGI | Kosovo Government Institutions |
| BI | Business Intelligence |
| DQA | Data Quality Assessment |
| DQI | Data Quality Improvement |
| DAMA | Data Management Association |
| DQI | Data Quality Indicators |
| DQD | Data Quality Dimensions |
| TDQM | Total Data Quality Management |
| DC | Data Cleansing |
| DL | Damerau Levenshtein |
| LA | Levenshtein Algorithm |
| ILA | Improved Levenshtein Algorithm |
| OCR | Optical Character Recognition |
| GSQA | Government Service Quality Assessment |
| AIMQ | A Methodology for Information Quality Assessment |
| CDQ | Comprehensive Methodology for Data Quality Management |
| COLDQ | Cost-effect of Low Data Quality |
| DQA | Data Quality Assessment |
| DQAF | Data Quality Assessment Framework |
| DQPA | A Data Quality Practical Approach |
| HIQM | Hybrid Information Quality Management |
| OODA DQ | The Observe-Orient-Decide-Act Methodology for Data Quality |
| TBDQ | Task-Based Data Quality Method |
| TIQM | Total Information Quality Management |
| WHO | World Health Organization |
| JHU | John Hopkins University |

| | |
|---|---|
| ECDC | European Centre for Disease Prevention and Control |
| NIPHK | National Institute of Public Health of Kosovo |
| GRNM | Government of the Republic of North Macedonia |
| G2C | Government to Citizens |
| PDR | Personal Documents Register |
| VR | Vehicle Register |
| DDD | Domain Driven Design |
| MVC | Model View Controller |

# 1 Introduction

## 1.1. Motivation

Nowadays, both practitioners and researchers of data are aware of how crucial it is to achieve and maintain a high standard of data quality. The impact of high data quality level is frequently considered as critical and valuable asset in development of enterprises.

High data quality has become a crucial component of data management within a business institution or organization. Since the beginning of the twenty-first century, there have been numerous notable technological advancements in the information technology sector, including cloud computing, the Internet of Things, and social networking. The advancement of these technologies has caused the rise of the volume of data in an exponential way [1].

Data migrations from outdated platforms like Excel, Access, etc. to more advanced relational database systems like MS SQL Server, Oracle, etc. have also contributed to an inadequate level of data quality. Also in this process has greatly influenced the non-condition of data types stored in certain fields where most of the fields have been text type.

Any sizable real-world dataset's quality is influenced by a variety of factors, but the most important ones that are frequently mentioned are the quality and the source of the data. Every stage of the process, from initial data gathering through archival storage, has the potential to introduce rogue or dirty data. According to research on several rogue data types, some of them are presented during the data entry process. [2].

In order to provide electronic services for public administration, a huge number of databases were created. However, because these databases were not connected or their data was not standardized, this resulted in a high number of mistakes and inaccuracies, which decreased the quality of the data.

For e-government institutions, evaluating and enhancing the quality of data held in information systems is a crucial and challenging process because the services offered directly depend on the quality of the data that is available [3].

Using various dimensions and approaches to improve the quality of data is a necessity for achieving high-quality electronic services.

The qualitative approach that will be used in this thesis to assess the quality of the data will be based on key dimensions such as [4]:

- Completeness
- Uniqueness
- Timeliness
- Accuracy
- Consistency.

There are several models and techniques for improving data quality and providing better e-services meaning that there are some challenges and open issues for research.

## 1.2. Research Problem

Nowadays in government institutions, the primary goal is to provide qualitative e-services for citizens. To provide qualitative e-services, a prerequisite is to have qualitative data in all government institutions.

Based on the literature review, case studies and various technical reports, the most common causes of producing poor data remain to be the human factor.

There are numerous current algorithms for data matching and connecting records between various data sources so it is very important to decide which algorithm is better to use for assessing quality and performance of data during treatment of the datasets.

We will concentrate on methods that handle enormous amounts of data, such as the Damerau-Levenshtein distance (DL) algorithm and the Levenshtein distance (LV) algorithm.

Inadequate data quality has also resulted from a number of issues, including:

- integration of data from multiple and different sources
- technical issues or faults
- using the volume of data for new proposes
- data entered manually
- software updates
- shortage of testing time.

Our goal is to analyze methodologies by using effective models, which will increase data quality in Government Institutions of the Republic of Kosovo with the reason for offering better e-services.

As well, the focus will also be on business intelligence (BI) because it is very important to incorporate data quality issues into all data integration and business intelligence processes from data sourcing to information consumption by the business user.

Data quality issues must be found as early as possible in the processes and also handled in accordance with the institution or business needs [5].

## 1.3. Hypothesis

The hypothesis arising from research questions are:

I. Harmonization process of the data will improve data quality outputs

II. Cleansing of a complex source data in Data Quality Services (DQS) will improve the data quality in e-services

III. The proposed model of data quality improvement positively influences the quality of data output.

IV. Data quality will increase with the use of suitable algorithms for comparing the techniques chosen for matching and connecting big datasets of personal records from various sources.

## 1.4. Research Questions

During my PhD work, I will try to answer the following issues:

1. What indicators might affect the quality of the data?
2. What are the root causes of poor data quality?
3. What are critical dimensions that need to be managed in data quality improvement?
4. What is the impact of migration of the data from different and old platforms to new platform in terms of data quality?
5. What is the impact of automatic interconnection of the different databases in terms of data quality?
6. What is the impact of implementing Business Intelligence in terms of data quality?
7. What is an appropriate algorithm for linking and matching large volumes of datasets to increase the quality of data and to provide better e-services?

All formulated research questions will address the situation with focus in Kosovo Government Institutions (KGI), so the datasets and other published data that will be used on this dissertation will be related to KGI. The datasets and results that will be used, will be in accordance with KGI under personal data protection law of Kosovo.

## 1.5. Research Methodology

The major objective of this thesis is to put into practice an approach for improving data quality so that it can be used, re-used, and redistributed by everyone. There will be

addressed the problems and issues as well faced during data quality improvement by defining the root causes of poor data quality, indicators that might affect data quality, impact of automatic interconnection and impact of implementing Business Intelligence (BI).

This dissertation's thesis approach will be based on a concept for designing science research procedures for information systems. [6].

In the following schema (Figure 1), it is shown the research methodology that will be applied to this thesis.

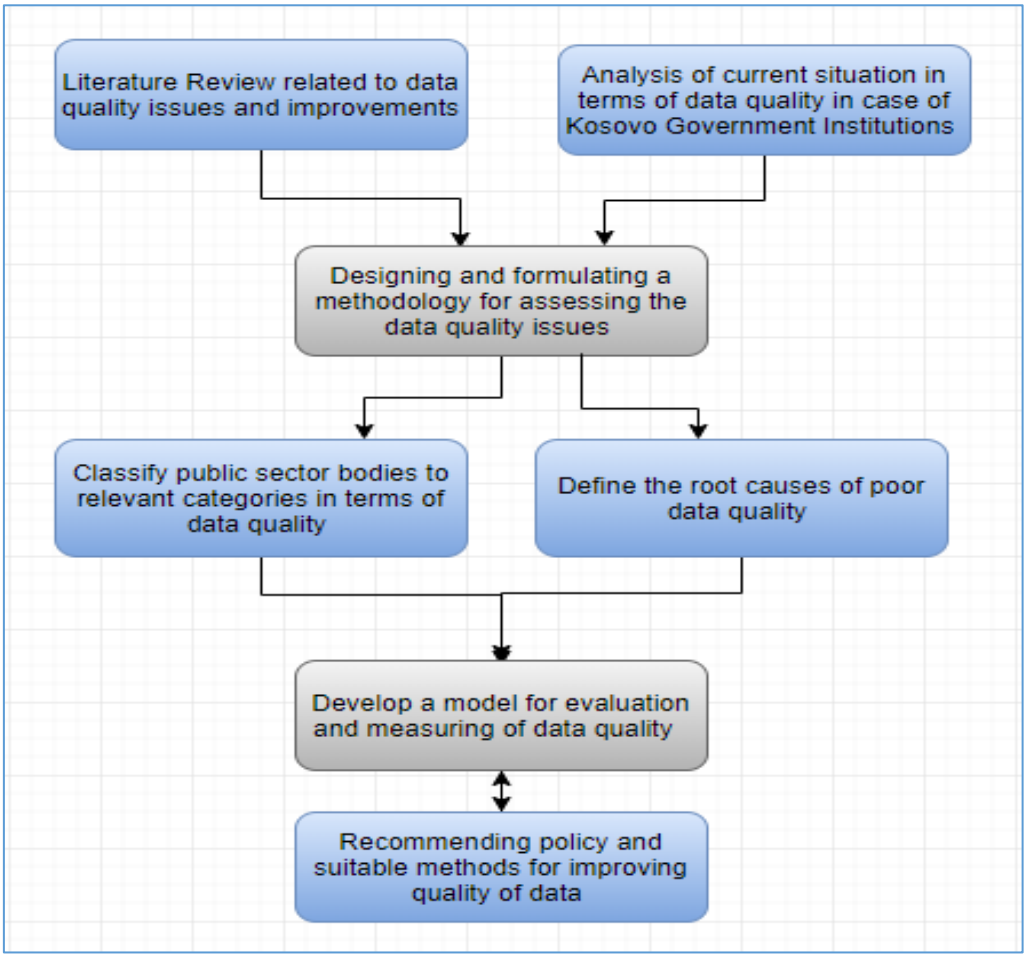



Figure 1: Research approach applied in the thesis.

The process of defining the attributes, dimensions, and metrics to analyze data is a fundamental activity in all methods for DQ assessment and improvement.

We will utilize a variety of specific and dedicated tools to assure data integrity and quality through data profiling, matching, cleansing, correcting, and monitoring the overall state of

the data cleansing process. One of the tools that will be used is SSIS Data Quality Services (DQS) [7], which is shown in the following Schema (Figure 2).

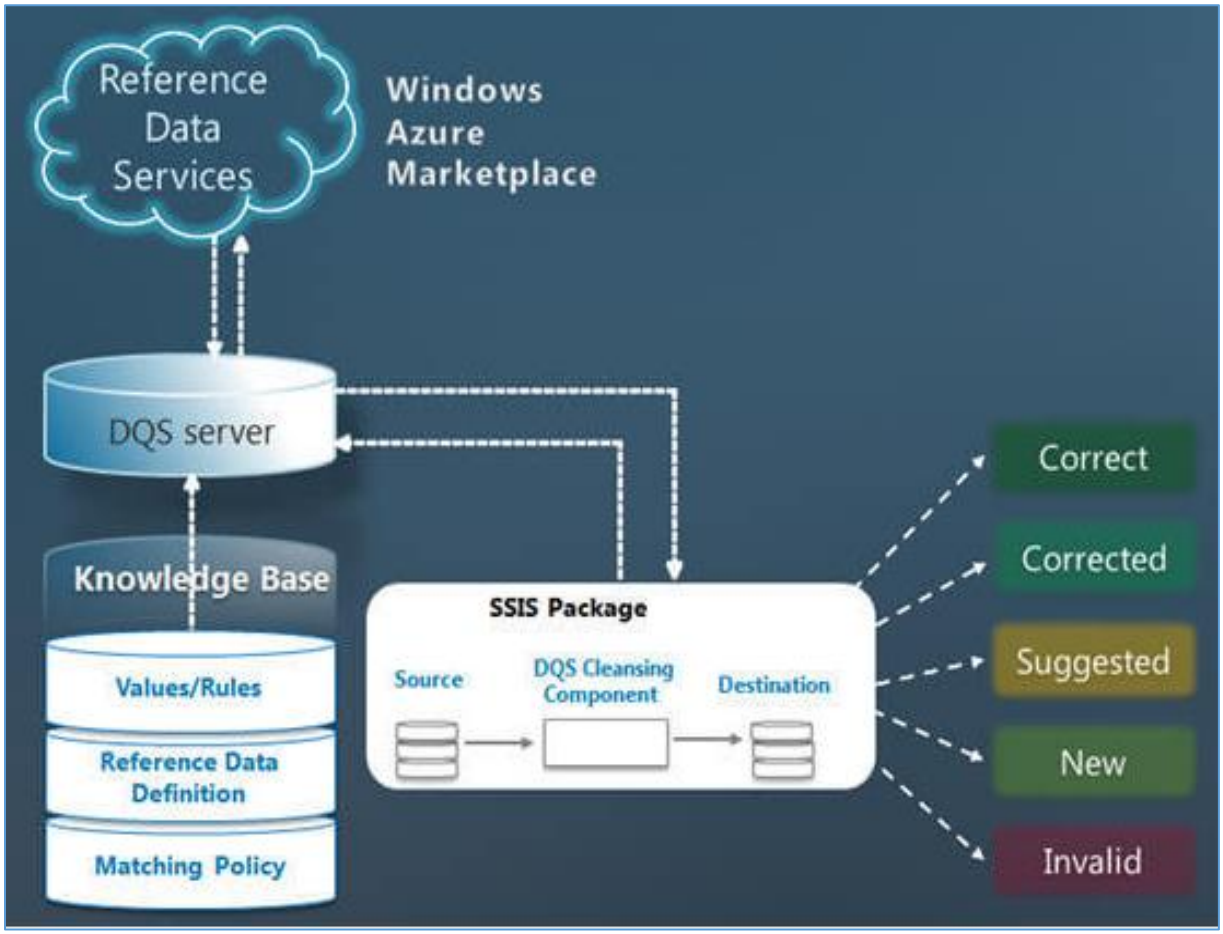



Figure 2: DQS Server for data cleansing and correction.

## 1.6. Qualitative and Quantitative Research

In this thesis we will combine methods with focus on both qualitative (dataset of n records) and quantitative (repeating process with improving dimensions each time) method.

## 1.7. Data Collection

Data collection will be done through collecting data from multiple resources and from different platforms and locations. All data that will be collected will be structural data from relational databases.

## 1.8. Data Analysis

The data will be analyzed using statistical tools and techniques in order to generate the results and conclusions. Graphical presentation will also be implemented in this thesis in order to obtain an illustrative presentation.

## 1.9. Publications

The research papers that have been published during this phase are:

1. Genc Hamzaj, Zamir Dika and Isak Shabani. “A Study on Comparative Analysis of COVID-19 Datasets”, Journal of Sciendo Review Volume 15 Issue 1: Challenges and Perspectives of Covid - 19, DOI: 10.2478/seeur-2020-0007 Sciendo, 2020.
2. Genc Hamzaj, Zamir Dika and Goce Armenski. “An Overview on How to Choose the Data Quality Assessment and Improvement Frameworks”, Proceedings of the 13th ICT Innovations Conference 2021, pp. 55-66.
3. Genc Hamzaj, Zamir Dika and Goce Armenski. “Comparison of the selected techniques for matching and linking large datasets of personal records from multiple resources”, Proceedings of the 13th ICT Innovations Conference 2021, pp. 158-169.
4. Genc Hamzaj, Zamir Dika, Isak Shabani. “E-Services Evaluation and Delivery Model Using Data Cleansing Logical Constraints – The Case of Kosovo Portal”, International Journal of Computer Engineering and Sciences Research, VOL. 04, NO. 02, May-June 2022, Pages 01–13 (ISSN: 2581-8481).

# 2
# Related Work

## 2.1 Introduction

This chapter analyzes the research work related to assessment of the impact of data quality for improvement of e-services in government institutions. Section 2.2 presents and discusses data quality dimensions used to determine the quality of data. Data quality indicators used for defining, measuring, analyzing and improving data quality are presented in Section 2.3. Section 2.4 treats problems and improvement of data quality. Section 2.5 assess frameworks for managing poor or dirty data. Section 2.6 presents methods for assessment and improvements of data quality. Section 2.7 presents algorithms for matching and linking records from multiple resources. Section 2.8 presents qualitative and user-friendly e-services through a government portal. Section 2.9 summarizes this chapter.

## 2.2 Data Quality Dimensions

Defining data dimensions is an essential component of assessing and improving data quality. A data component that can be measured or evaluated against a set of criteria to determine data quality is referred by specialists as a "Data Quality (DQ) dimension" [4].

From the literature, there are proposals for different sets of data quality dimensions and for which there is no clear definition of which set of data quality dimensions can be evaluated as "good" data dimensions, depending on the contextual nature of the data quality [2].

According to DAMA UK Working Group, there are six core dimensions of data quality as shown in figure 3 [4]:


Figure 3: Core dimensions for assessing quality of the data [4].

According to Scannapieco, M, & Catarci, T., the majority of authors consider the following dimensions for DQ: accuracy, completeness, consistency, timeliness, interpretability, and accessibility [8].

In general, works on the classification of data quality dimensions can be divided into two categories: academics view of data quality dimensions; and practitioners view of data quality dimensions [9].

According to the CDC, some of the strategic actions for Data Quality Assessment that should be considered are [10]:

1. Data quality assessment of data items that are determined based on importance for business operations
2. The use of data quality dimensions based on the determination of their weighting, where if they are ranked according to the ease of assessment, it starts with

completeness and validity, then continues with timeliness and uniqueness, while at the end it is done with accuracy and consistency as dimensions with a more difficult level to assess

3. Determine the value or levels that represent optimal and poor data quality for each dimension of data quality
4. The data items should be assessed using the criteria
5. Examine the results to see if the data quality is satisfactory
6. Make necessary corrections
7. To keep track of changes in data quality, repeat the previous steps periodically.

In table 1 are shown all dimensions defined in the ISO standard ISO/IEC 25012:2008 that has enacted in 2008 [11] for what in the standard is defined as DQ, that is "the level where the data meet the needs and inside a computer system for data quality it offers a generic model in a structured form for these data."

| **DQ characteristic** | **Definition** |
|---|---|
| Correctness | Represent the desired attribute's real value in the intended usage situation |
| Completeness | In a particular usage context, subject data has values for all anticipated attributes |
| Consistency | Are cohesive and free of conflict in the usage environment that they are intended for |
| Credibility | Believed to be true and credible by users |
| Accessibility | Citizens who require tech support or unique configuration due to a specific limitation can access data |
| Compliance | Comply with current standards, conventions, legislation, and other data quality guidelines |
| Confidentiality | Make sure that only permitted people may access and comprehend it |
| Efficiency | Could be accomplished by utilizing the right actions to deliver the desired levels of performance |
| Precision | Precise value when used in a particular context |
| Traceability | Offer an audit history of actions of who has accessed the data and who has modified data |
| Availability | Stated in the proper languages, symbols, and units, and are readable and understandable by users |
| Portability | Possibility of install, replace, or transfer from one system to another |

| | |
|---|---|
| | while maintaining the current quality |
| Recoverability | Allow it to continue operating and preserving a certain degree of quality even where is a defect |

Table 1: ISO standard DQ dimensions [11].

According to Jack Tan, the six main dimensions are presented in the figure below, which are mostly mentioned based on the study of 15 different data quality assessment methodologies where 32 dimensions were included [12].

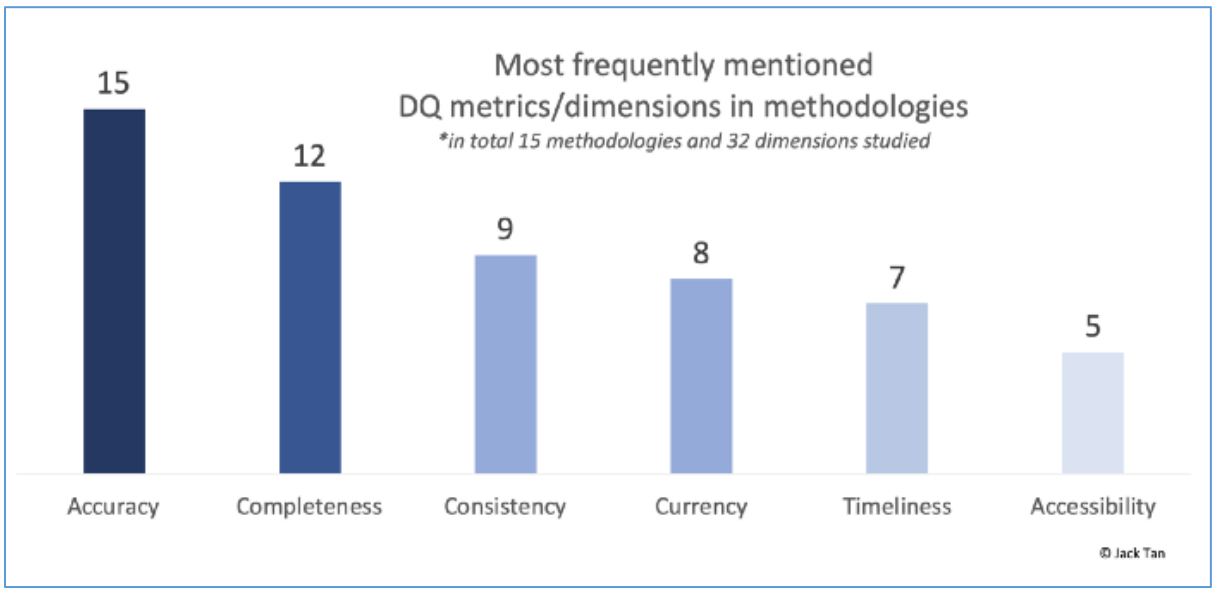



Figure 4: Six most frequently used data quality dimensions [12].

Depending on the selected dimensions, the assessment of data quality by making a plan can be based on some of the following requirements [12]:

- **Timeframe**
- **Definition**
- **Aggregation**
- **Interpretability**
- **Threshold**.

According to Le, Zhang and Shi, the table below shows an example of some dimensions based on data quality standard analysis with their explanations and examples [13]:

| Requirements | Explanation | Dirty data examples |
|---|---|---|
| Correctness | The data must reflect the true reality | Age=125 or input birthday= '12/12/1922' in fact do not know his birthday |
| Completeness | Datasets provide all the information required | Lack lost data of customers |
| Consistency | In different systems, there are codes that eliminate conflicts during interoperability. | One system has a customer's ID '1234, while in other system it is '087654 |
| Minimality | Each registration is unique and is not repeated during integration | After the integration we have several records for the same sale |
| Reliability | Regardless of the author, there is stability of the integration result | In the given table results the change of any attribute between the integration processes. |

Table 2: Analysis of data quality standards from a data mining perspective [13].

## 2.3 Data Quality Indicators (DQI)

Ensuring the highest quality data is achieved through continuous actions of measurement, analysis and improvement of data quality. In general, DQ assessment includes of numerous phases that an organization, users, and developers must perform, shown in Figure 5 [14]:

1. Within the required context, the definition phase determines relevant DQ dimensions.
2. The measurement process establishes and generates the metrics and measures required to evaluate DQ.
3. The analysis phase define the root causes of DQ issues and determines the impact of low-quality data.
4. In the improvement step, the most appropriate techniques are applied to improve DQ.

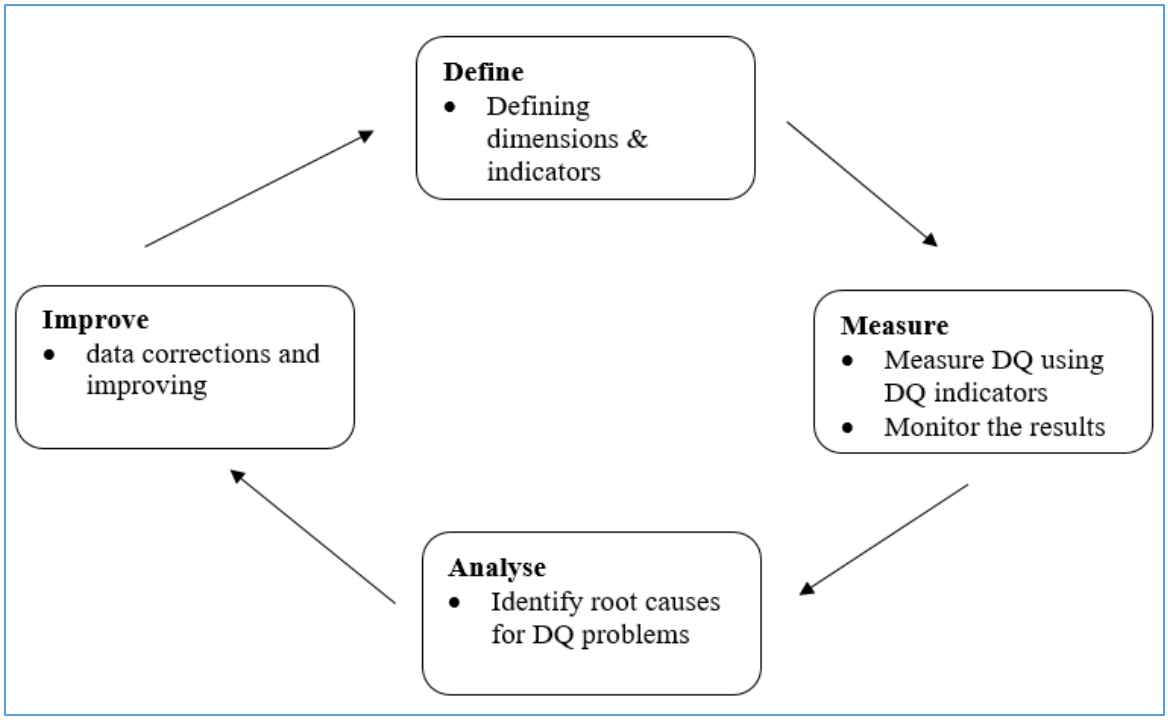



Figure 5: Process to insure high quality data [14].

In the following table are shown indicators for improving DQ [1].

| Dimensions | Elements | Indicators |
|---|---|---|
| | | |
| 1) Availability | 1) Accessibility | • Whether is a mechanism for data communication |
| | 2) Timeliness | • If we have communication with data within the specified time<br>• Whether data are regularly updated |
| 2) Usability | 1) Credibility | • The data comes reliably from the designated authorities<br>• The content of the data is correct based on regular checks by experts<br>• Data exist in the range of acceptable values |
| 3) Reliability | 1) Accuracy | • Data provided are accurate<br>• Data representation (or value) reflects the reality from the source |

| | | |
|---|---|---|
| | | information |
| | 2) Consistency | • The data with its attributes still match as before the processing<br>• Data are still consistent and verifiable |
| | 3) Integrity | • Data meets the criteria<br>• Data are consistent with structural and content integrity |
| | 4) Completeness | • Whether the set of data is complete and without missing any component |
| 4) Relevance | 1) Fitness | • The majority of data sets found match the users' requested themes |
| 5)Presentation Quality | 1) Readability | • Data is understandable thanks to factors like content, presentation, etc. |

Table 3: Dimensions, elements and indicators of assessment of DQ [1].

According to Hong Chen, there are two categories of data quality attributes: the good data quality category and the poor data quality category [15].

In the table below are shown grouped attributes.

| **Item** | **Attribute** |
|---|---|
| High data quality | Completeness, accuracy, timeliness, validity, relevance, reliability, integrity, confidentiality, comparability, consistency, usability, objectivity, importance, use of standards, accessibility, transparency, etc. |
| Poor data quality | Missing data, inconsistencies, data errors, invalid data, illegible handwriting, etc. |

Table 4: Attributes of DQ [15].

## 2.4 Problems and Improvement of DQ

DQ is frequently defined as acceptability of the data for use in different needed usage purposes, which must be error-free, comprehensive, accurate, current, and consistent. Data flaws, missing numbers, typos, inconsistencies, improper formatting, and other flaws in the data are all examples of poor data quality. During the processes of data cleansing and transformation, poor quality data must be corrected [16].

Some of the data quality problems are as following [16]:

- Missing values
- Incorrect information caused by input or processing errors
- Duplicates in the dataset
- Unevenly represented data
- Logically contradictory values.

According to Rahm and Do, the below figure is the categorization of DQ problems [17].

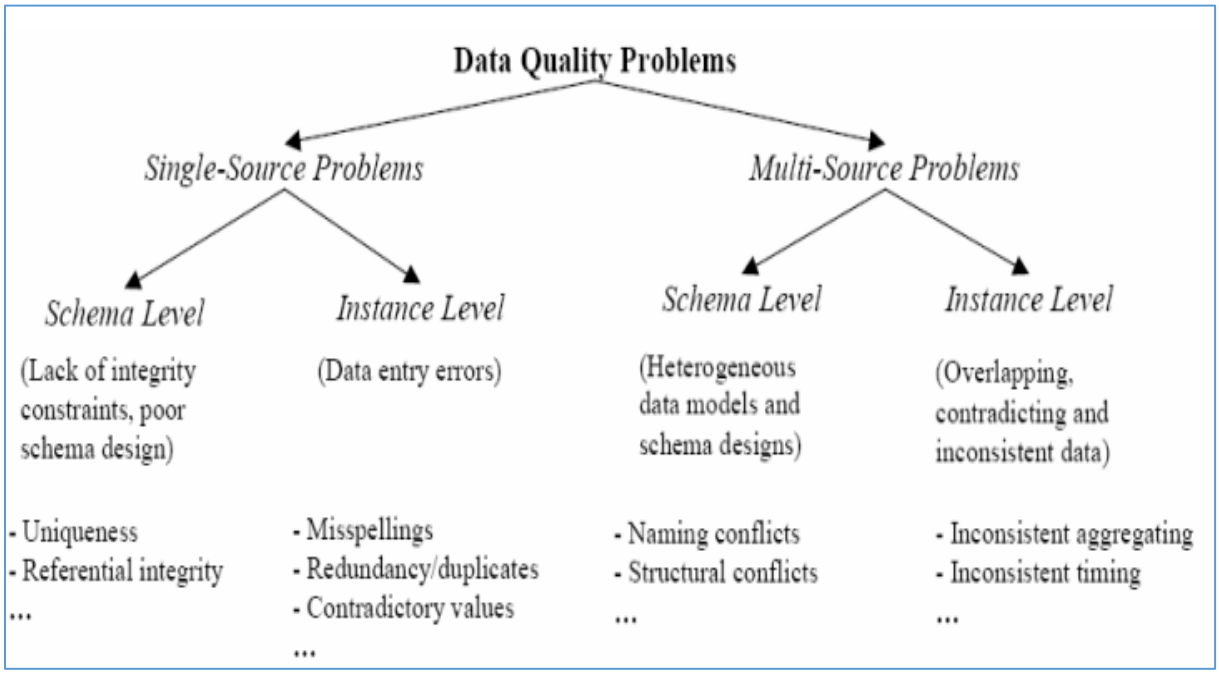



Figure 6: Categorization of DQ problems [17].

The process of detecting and correcting errors and anomalies in order to improve the quality of certain data sources is referred to as data cleansing (DC or "Data Cleaning") [18]. Multiple operations are done on low quality data, including processes for defining incomplete data, erroneous, out-of-date, inconsistent, redundant, etc., where these operations are accomplished throughout this data cleansing utilizing a broad variety of specific methodologies and technologies. It subdivides them into the following phases (as shown in Figure 7) [16]:

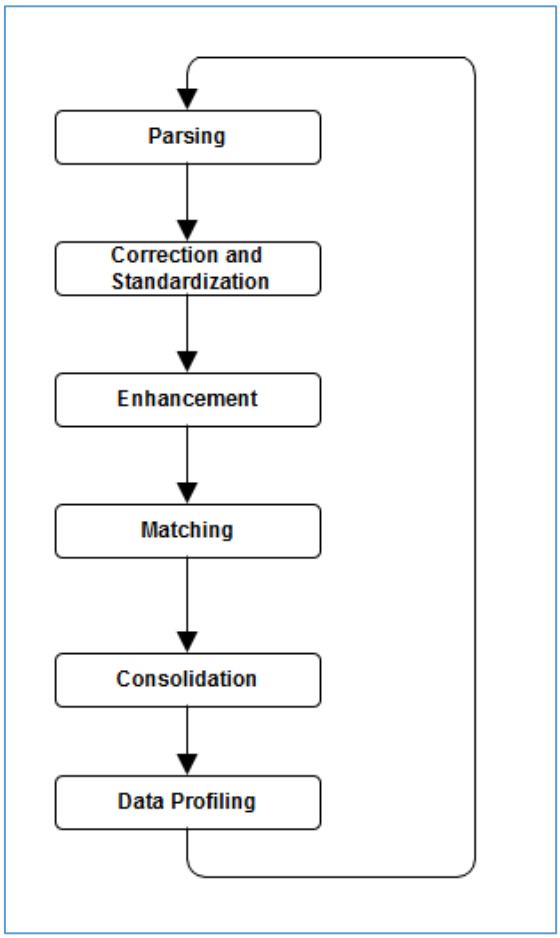



Figure 7: Data Cleaning Process [16].

Figure 8 provides a visual representation of the process flow [16]:

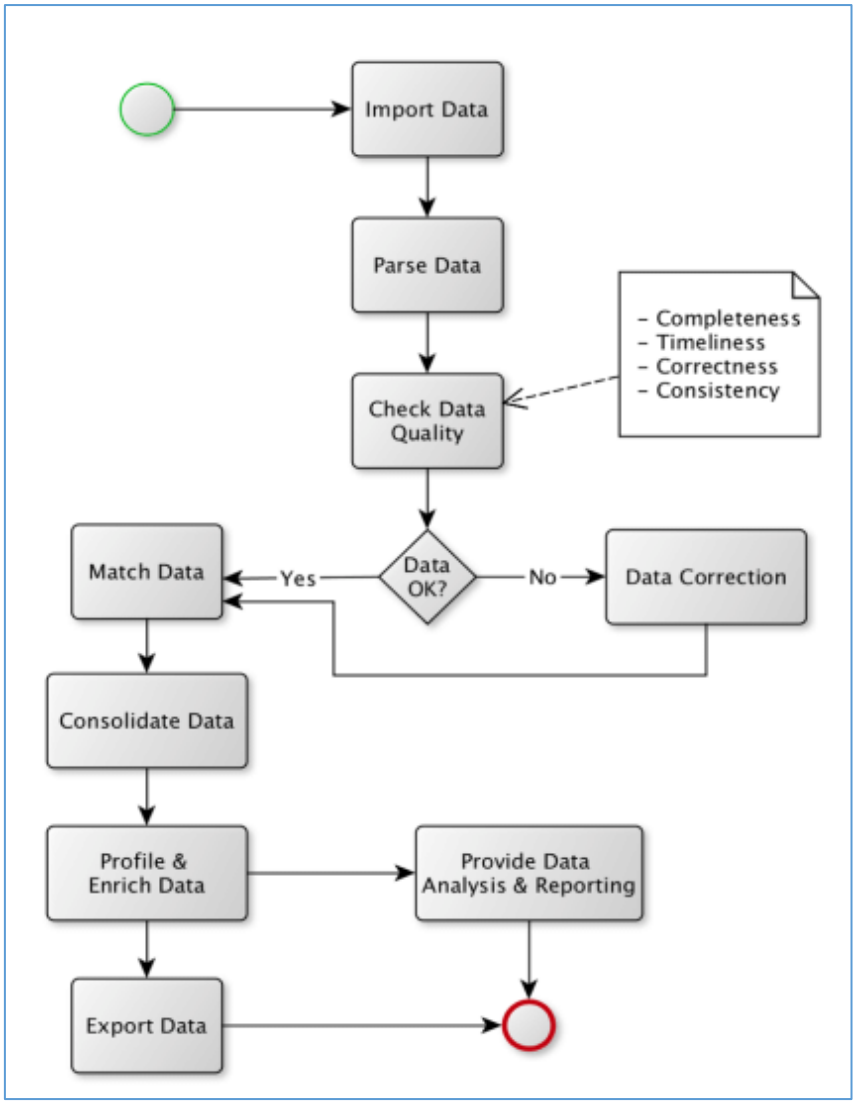



Figure 8: Main Workflow of the Process [16].

## 2.5 Frameworks for Managing Poor or Dirty Data

According to Tony O'Brien, the data governance model should be built around three interconnected essential elements: People, Processes and Data, where any attempt to improve the quality of data within any organization must be focused around these three essential elements. Conceptual Framework is shown in Figure 9 [19].

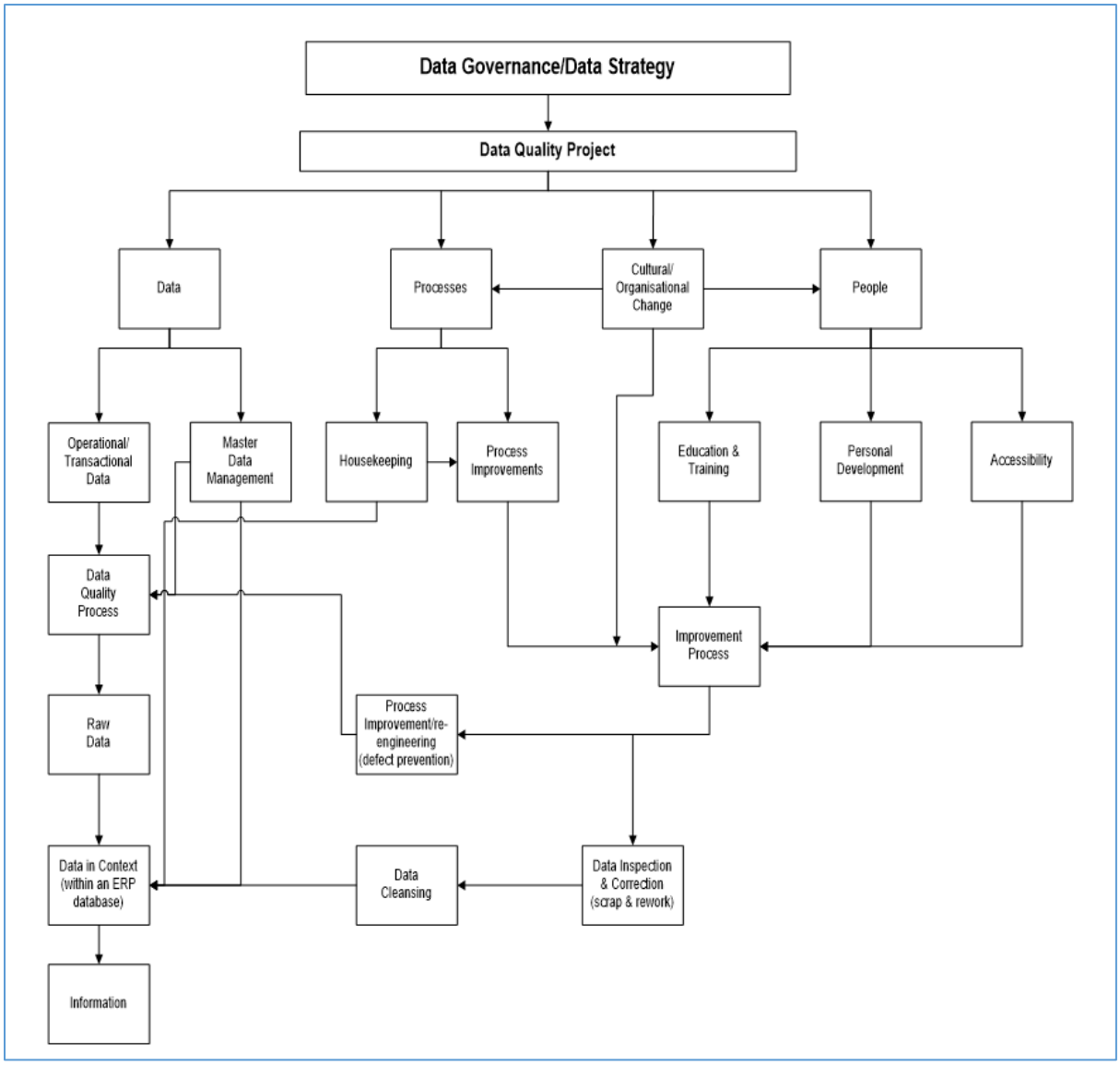



Figure 9: Conceptual Framework [19].

In this age of big data, with the growth of data sources types, data users are not always data producers. Therefore, Li Cai, from the view of the users, propose a multilayer data quality standard, as shown in Figure 10 [1].

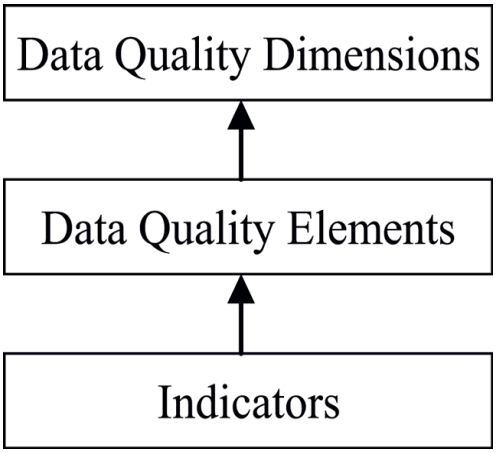



Figure 10: Data quality framework [4].

An appropriate data quality assessment process based on the properties of big data is shown in Figure 11 [1].

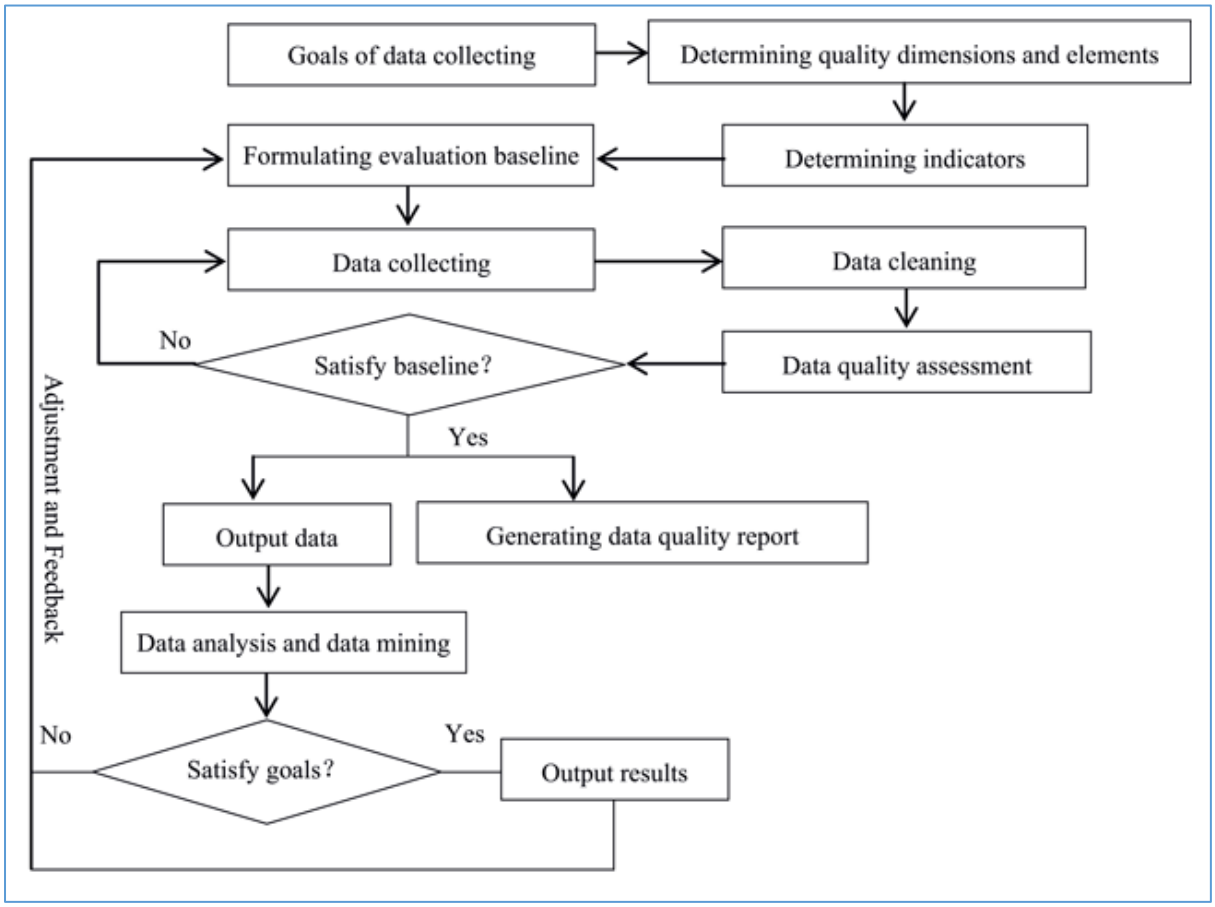



Figure 11: Quality assessment process for big data [1].

## 2.6 Methodologies for Assessment and Improvement of Data Quality

According to Woodall, P., Oberhofer, M., & Borek A., without the use of proper information system approaches, such are algorithms that computers can run automatically to find and correct possible dataset errors, data quality assessment and improvement are typically not achievable [20].

According to Woodall, P., Borek, A., and Parlikad, A, the goal is that through examining the data to detect the existing DQ level [21].
DQ Methods for assessment of data quality problems are [20]:

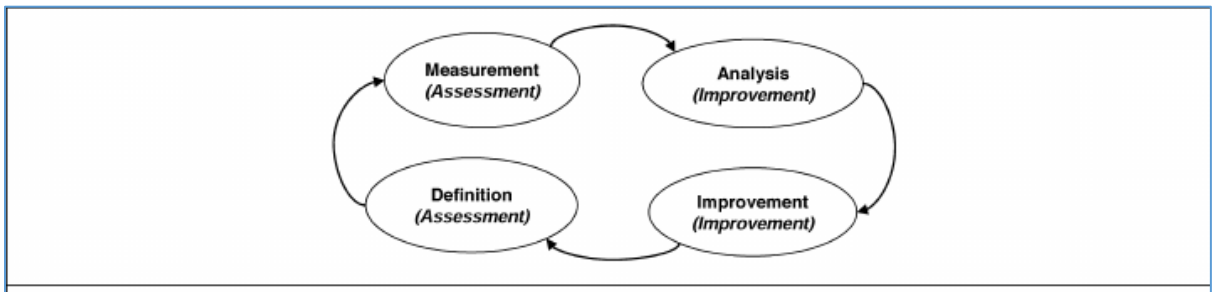
- Column analysis
- Cross-domain analysis
- Data verification
- Domain analysis
- Lexical analysis
- Matching algorithms
- Primary key and foreign key analysis
- Schema matching
- Semantic profiling.

After the detection of the current problems from the DQ assessment methods, there are improvement methods, which have as their final goal the correction of the data. Some of the DQ improvement methodologies that can be used for data correction are listed below [20]:

- Data standardization
- Data enrichment
- Data consolidation
- Data integration.

According to Wang, the TDQM (Total Data Quality Management) methodology in the literature is presented as one of the first general methodologies ever published [22]. Although TDQM is presented as a discovery of academic research, it is also widely used as a guide in organizational data reengineering initiatives. TDQM's main goal is to further expand the principles of Total Quality Management in the aspect of data quality.

Phases of TDQM are listed in figure below.




Phase 1: **Definition**
- Input → Information product
- **Data Analysis**: Definition of the characteristics at two levels: (1) high level: description of functionalities for information consumers (2) low level description of the basic units and components of the information product and their relationships
- **DQ Requirements Analysis**: Definition of the IQ requirements from the perspective of IP suppliers, manufacturers, consumers and managers.
- **Process Modeling**: Definition of the Information Manufacturing System
- Output → Logical and physical design of the Information Product with the necessary quality attributes.
  - A quality entity-relationship model that defines the IP and its IQ requirements
  - An information manufacturing system that describes how the IP has been produced

Phase 2: **Measurement**
- Input→ IQ dimensions
- **Measurement of quality**: Definition of the IQ metrics
- Output → IQ problems

Phase 3: **Analysis**
- Input → IQ problems
- **Identification of the causes of errors**: Determination root causes of discrepancies
- Output → Actions for improving data quality

Phase 4: **Improvement**
- Input → IQ metrics
- **Selection of strategies and techniques**: Identification of key areas for improvement
- Output → IQ improvement techniques

Figure 12: Phases of TDQM [57].

According to LEE, Y.W., STRONG, D. M., KAHN, B. K., ANDWANG, R. Y., AIMQ (A Methodology for Information Quality Assessment) Methodology focuses primarily on benchmarking as an independent and objective technique for assessment data quality [9].

Phases of AIMQ are listed in figure below.

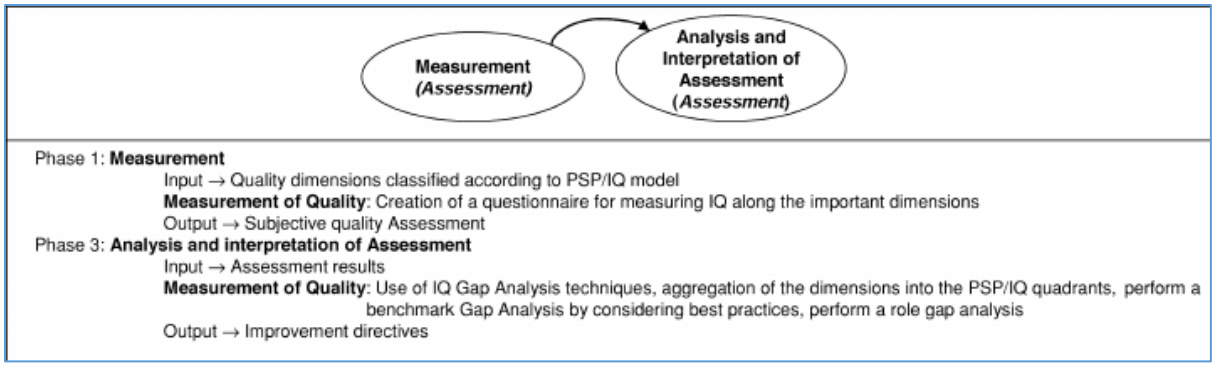



Figure 13: Phases of AIMQ [57].

## 2.7 Algorithms for Matching and Linking Records from Multiple Resources

In order to perform data matching and linking from different sources, different existing algorithms can be used, which greatly facilitate this process. The data collected from different sources often do not have good quality, so the intention is to improve this data with the main aim of providing better e-services.

Some of these algorithms are:

- Levenshtein distance algorithm
- Damerau Levenshtein distance algorithm
- Optimal String Alignment
- Q – gram distance
- Longest Common Substring
- Jaccard distance Cosine distance.

A. LEVENSHTEIN DISTANCE ALGORITHM

The Levenshtein distance is an algorithm used to measure the difference between two given sequences. Informally, the Levenshtein distance is the minimum number of operations or modifications (e.g. Insertion, Deletion or Substitution) required for a single-character of the first word until it will be the same as the second word [23].

According to Nikhil Babar, through the Levenshtein algorithm it is possible to determine the least amount of operations required to change one string and turn this string into another. It can be calculated effectively using below approach. [23]:

- In order to initialize a matrix, the (m, n) cell's distance between a word's m- and n-character prefixes must be determined
- The upper left to bottom right corners of the matrix can be filled in
- An insert or a deletion is represented by each hop, whether it is horizontal or vertical
- Normally, the cost for each operation is set to 1
- If both characters in the row and column match, it will be either one or zero. Every cell always attempts to reduce local costs
- In this situation, the Levenshtein distance between the two words is represented by the number in the lower right corner.

According to Rishin Haldar and Debajyoti Mukhopadhyay, following steps must be taken by the algorithm [24]:

| **The Levenshtein Distance Algorithm** | |
|---|---|
| **1:** | **Step 1: Initialization** |
| **2:** | a) Set n to be the length of s, set m to be the length of t |
| **3:** | b) Construct a matrix containing 0..m rows and 0..n columns |
| **4:** | c) Initialize the first row to 0..n |
| **5:** | d) Initialize the first column to 0..m |
| **6:** | **Step 2: Processing** |
| **7:** | a) Examine s (i from 1 to n) |
| **8:** | b) Examine t (j from 1 to m) |
| **9:** | c) If s[i] equals t[j], the cost is 0 |
| **10:** | d) If s[i] doesn't equal t[j], the cost is 1 |
| **11:** | e) Set cell d[i,j] of the matrix equal to the minimum of: |
| **12:** | i) The cell immediately above plus 1: d[i-1,j] + 1 |
| **13:** | ii) The cell immediately to the left plus 1: d[i,j-1] + 1 |

**14:** iii) The cell diagonally above and to the left plus the cost:

**15:** d[i-1,j-1] + cost

**16:** **Step 3: Result**

**17:** **Step 2 is repeated till the d[n,m] value is found**

---

B. DAMERAU LEVENSHTEIN DISTANCE ALGORITHM

The Damerau Levenshtein distance represents a variant of another form of the Levenshtein distance, where it pertains to algorithms of the Edit type. As was previously mentioned, the "edit-distance" category determines how distinct two strings are by turning one string to the other and calculating the operations needed.

The Damerau-Levenshtein distance, compared to the classic Levenshtein distance, during the character edit, in addition to operations such as Insert, Delete and Substitution, also uses the transposition operation [25].

Wagner and Fischer [26] created a trace notion as a function of cost in several structures, in order to simplify the process of finding the distance between the first string and the second string.

This trace can be illustrated as a diagram as in the figure below.

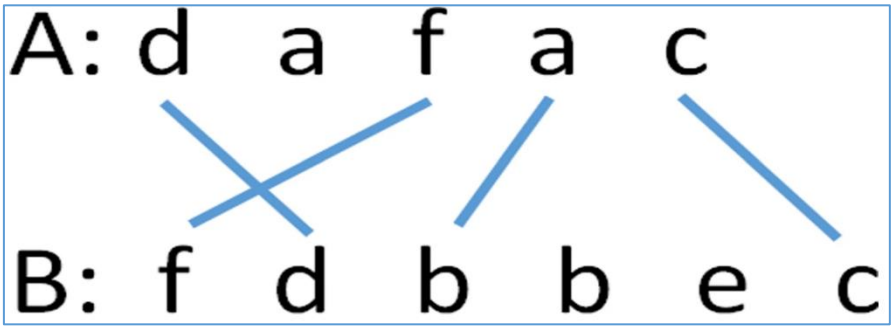



Figure 14: DL trace example [26].

Below is the pseudocode of the Damerau Levenshtein algorithm, where the H value is calculated, whereas last_row_id[c] represents the last trace of character c in A and last_col_id represents the last trace of $a_i$ in B [27].

**Damerau Levenshtein Distance Algorithm**

```
1:  DL(A[1:m], B[1:n])
2:  for j ← 0 to n do
3:      H[-1] [j] ← maxVal; H[0] [j] ← j
4:  end for
5:  for i ← 1 to m do
6:      H[i] [-1] ← maxVal; H[i] [0] ← i
7:      last_col_id ← -1
8:      for j ← 1 to n do
9:          diag ← H[i -1] [j -1] + c(A[i], B[j])
10:         left ← H[i] [j-1]+1
11:         up ← H[i-1] [j] +1
12:         k = last_row_id[B[j]], l = last_col_id
13:         transpose ← H[k-1] [l-1] + (i – k - 1) +1 + (j - l -1)
14:         H[i] [j] ← min{diag, left, up, transpose}
15:         if A[i] = B[j] then
16:             last_col_id ← j
17:         end if
18:     end for
19:     last_row_id[A[i]] ← i
20: end for
21: return H[m] [n]
```

### 2.7.1 Improvement of Algorithms for Matching and Linking Records from Multiple Resources

Many researchers used different ways and methods with the aim of improving algorithms for matching and linking records from multiple resources.

According to H.N. Abdulkhudhur & I.Q. Habeeb, Levenshtein's algorithm is the most used algorithm for finding words that are most similar to the incorrect word based on a certain

lexicon. By sequentially contrasting the letters of the incorrect word with the characters of the correct word from a lexicon, it calculates a sequence of operations that fill the cells of an array. Such actions will be done millions of times for each wrong word to create the list of viable options. In order to reduce this large number of operations created due to the comparison of the characters of the incorrect word and the lexicon words, the authors propose an improved Levenshtein algorithm. Compared to Levenshtein's algorithm, the proposed so-called ILA-OT algorithm, based on experimental results, has a reduction in processing time of 32.43% [80].

From Figure 15, it can be seen that in terms of processing time, the proposed ILA-OT algorithm is faster than the LA algorithm in percentage by about 32.43%, while not changing the number of comparisons between both algorithms with the total number of the comparisons as shown in Figure 16, with 100% accuracy [80].

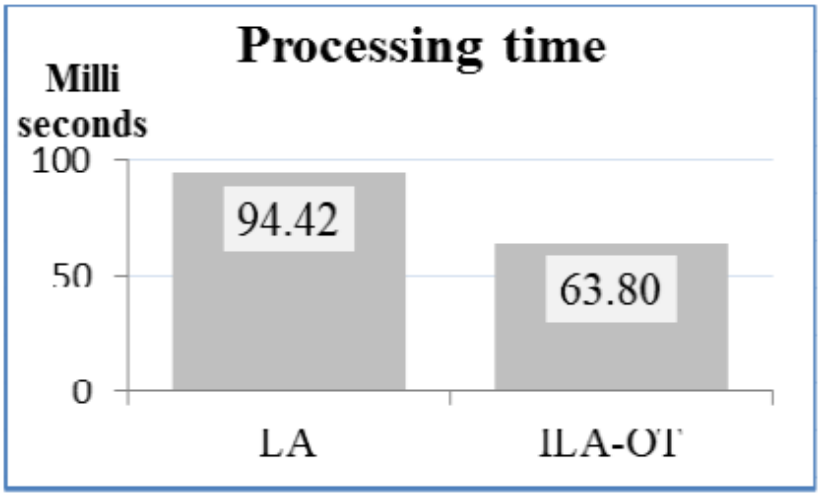



Figure 15: Processing Time LA and ILA-OT [80].

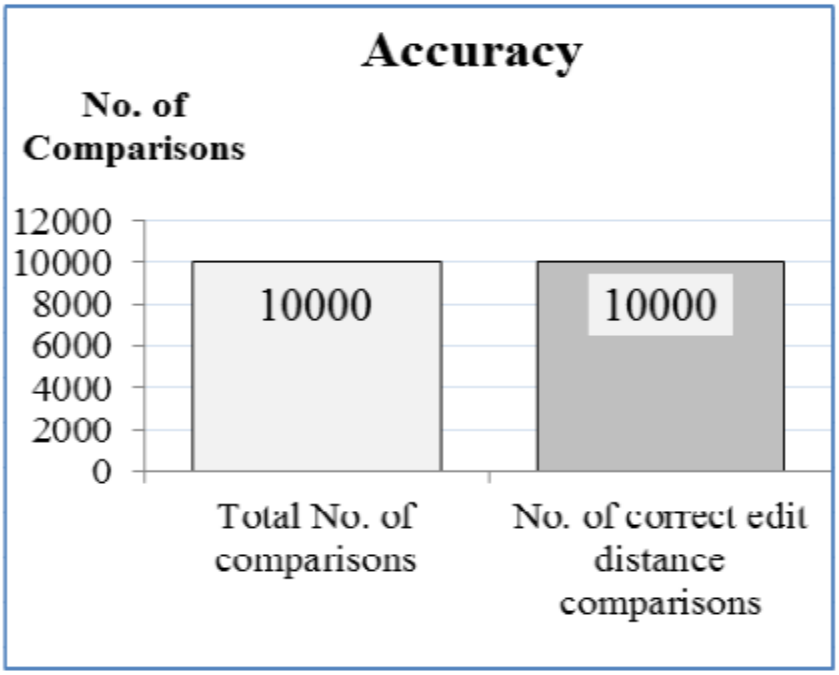



Figure 16: Accuracy between LA and ILA-OT [80].

According to Z. ZHAO & Zh. YIN, extending the transposition procedure in the current method reduces the number of edit operations. Improving the algorithm in such a way that by enabling transposing isolated symbols even after the calculation position and not only before the calculation position, then as a result a better edit distance can be obtained [81].

According to Shama Rani & Jaiteg Singh, by removing stop words such as “also, is, am, are, they, them, their, was, were” etc., Levenshtein's edit distance algorithm can be modified and improved [82].

Due to their frequent occurrence in the language, these words are designed by the majority of search engines to be disregarded while indexing or retrieving the result from the search engine [4].

The following conditions lead to the removal of stop words [82]:

- Each manuscript has around 20–25% stop words
- Eliminating stop words increases the effectiveness of the document
- Text mining and searches do not benefit from stop words
- Aim is to reduce indexing.

Levenshtein's Edit distance algorithm is utilized to calculate the inputs and the amount of words in all the manuscripts, as shown in table 5. The time taken to calculate Levenshtein's distance with Stop words is shown in table 6. The time taken to compare documents is calculated in milliseconds [82].

| Text Length of Document A | Text Length of Document B | Document A after removing stop words | Document B after removing stop words |
|---|---|---|---|
| 51 | 62 | 27 | 38 |
| 103 | 90 | 59 | 53 |
| 203 | 192 | 124 | 119 |
| 395 | 410 | 242 | 233 |
| 798 | 750 | 470 | 474 |

Table 5: Text length of Document A and B with and without using stop words [82].

| Text Length of Document A | Text Length of Document B | Time taken to calculate Levenshtein's distance with Stop words (in milliseconds) |
|---|---|---|
| 51 | 62 | 14 |
| 103 | 90 | 16 |
| 203 | 192 | 23 |
| 395 | 410 | 62 |
| 798 | 750 | 218 |

Table 6: Length of time needed to determine LV distance once stop words are removed [82].

Figure 17 compares the lengths of Document A and the identical document after the stop words were deleted. Additionally, Figure 17 shows the text length with and without stop-words [82].

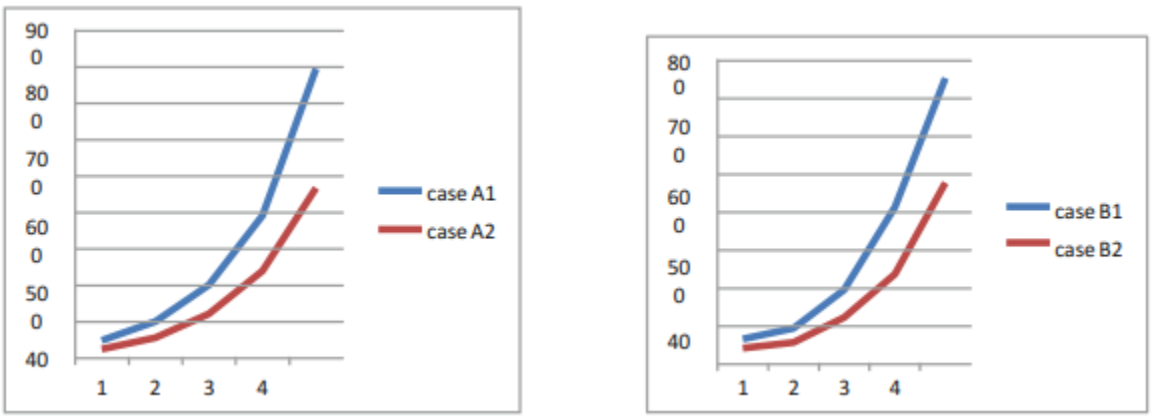



Figure 17: Document A and B by using or avoiding stop-words [82].

Figure 18 shows the time taken to calculate LA with stop-words (case TW1) and the time needed to compute edit distance following the removal of stop words (case TW0) [82].

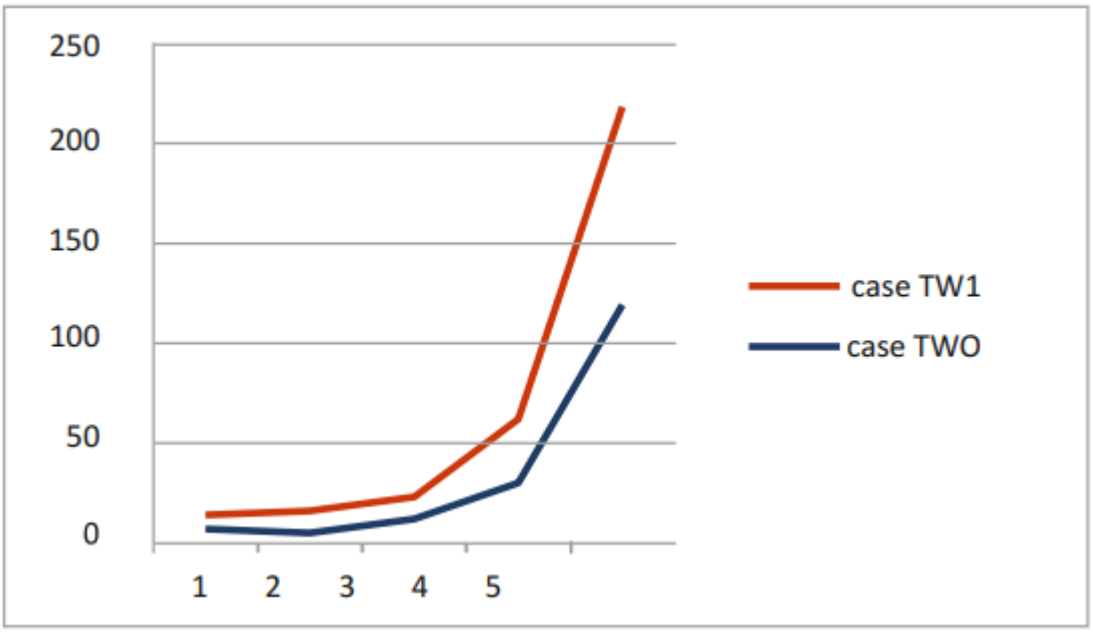



Figure 18: Time spent on calculations both before and after stop words were removed [82].

According to R. Haldar and D. Mukhopadhyay, in cases where the letters are not recognized by Optical Character Readers, dictionary lookup methods are mostly used. However, these methods increase the cost of searching due to the complexity in the calculation, so the Levenshtein distance is an effective algorithm with the aim of string approximation.

As is known, the Levenshtein Distance Algorithm, for any operation (Insert, Delete, or Substitute) gives the uniform distance value (ie, 1) when comparing two different characters. An improvement of this method by grouping characters with similar appearance and calculating the difference of the characters of this group with a value smaller than the value 1, as a result would enable closest matches to be more accurate. For example, during the use of any operation for the characters O, D, Q, the weight may be given with a value of 0.4 and not 1 as for the other characters. As a result, we will have an improved version compared to the initial version of Levenshtein's algorithm [103].

The groups identified are [103]:

1) O, D, Q

2) I, J, L, T

3) U, V

4) F, P

5) C, G

For example, it may happen that OCR has mistakenly read the word BODY as BDQY due to the similarity of the characters O and D. Figure 19 explains this case quite clearly. From the word BDQY, according to Levenshtein's algorithm, all other words (BODY, BUSY, BURY, BONY) have a distance of two. However, if we use the improved version of Levenshtein Distance Algorithm, it turns out that the distance between BDQY and BODY has the shortest distance compared to the other words, because D, Q is in the same group as O, D, so the word BODY is chosen as the answer [103].

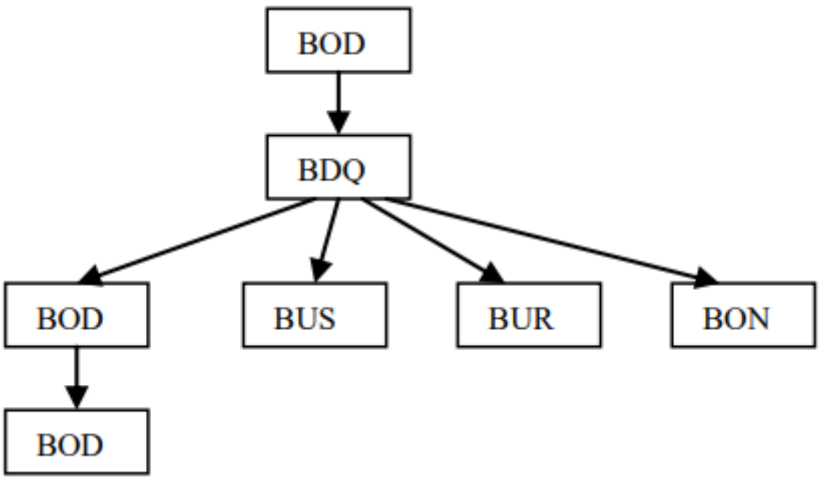



Figure 19: An example of possible outcomes. [103].

## 2.8 Qualitative and User Friendly e-Services Delivery through Government Portals

When developing e-services in government institutions, the focus is to offer qualitative e-services with less complexity as possible. Before offering e-services, assessment and improvement of quality of data for the services provided is very important because it affects the trust of citizens for the services delivered.

The rapid development of information and communication technology has altered how governments and their citizens communicate, and as a result, a new type of governance known as electronic government has emerged [28]. Understanding the relationships between the effectiveness of government services and its effects on the pleasure of users and citizens has become critically important. Dimensions of data quality have become a crucial resource for e-governments to assess and evaluate electronic platforms, user satisfaction and to provide better e-services.

According to Taiseera Hazeem AlBalushi, in order to ensure the happiness and loyalty of users and citizens, it will become increasingly difficult for the electronic government in the future to maintain the quality of electronic services. It is also crucial to evaluate the service's quality using the dimensions, sometimes referred to as quality factors or indicators with the aim to achieve this goal. The most important quality dimensions of data are grouped in Table 7 with specific focus (service providers, service users) [29].

| Dimension | Model/instrument | Focus<br>SU – Service Users<br>SP – Service Providers |
|---|---|---|
| Personalization | The level or degree of modification that consumers can manage based on their wants and requirements is referred to as personalization | SU |
| Usability | Usability evaluates a user's efficacy, efficiency, and contentment using e-government portals.<br>The availability of all relevant information, which is additionally helped by sophisticated choices for searching for crucial information, is used to | SU |

| | | |
|---|---|---|
| | explain why e-government portals are simple to use | |
| Performance | Through Performance we are able to evaluate the dependability and timeliness of the e-government services we have used | SU |
| Web design | As the single point of contact between the citizen portal and e-government, the dimension that might be also referred to and considered as a point of contact. Dimension relates to the entire design of the website, which includes the accessible information as well as the design | SU and SP |
| Security | This variable measures users' impressions of the safety and security of their personal information when using an online portal.<br><br>Citizens are more likely to use e-government portals for their needs when they feel confident knowing that whatever services they use are completely safe and secure. | SU and SP |
| User involvement | The dimension includes communication between citizens and the electronic portal, which is done through feedback on the services offered and notifications of service failures via email, SMS, and other electronic channels. | SU |
| Satisfaction | The percentage to which a customer's wants and expectations are satisfied by the service provided is measured by their level of satisfaction. This | SU |

| | | |
|---|---|---|
| | particular component aims to gauge overall satisfaction with and favorable attitudes regarding utilizing e-government services. | |
| Loyalty | The loyalty component or dimension and the level of service satisfaction delivered by e-government portals are intimately correlated. Through high percentage of using the e-service through the government site, it is possible to show and measure loyalty | SU |

Table 7: E-services quality dimensions for achieving user satisfaction and loyalty [29].

According to Anjoga H., Nyeko S. and Kituyi M., these authors confirm that electronic government usability benefits government stakeholders by facilitating better governance and faster, more secure public access to information resources. Figure 20 shows the framework proposed for usability of e-Government services in developing countries [30]:

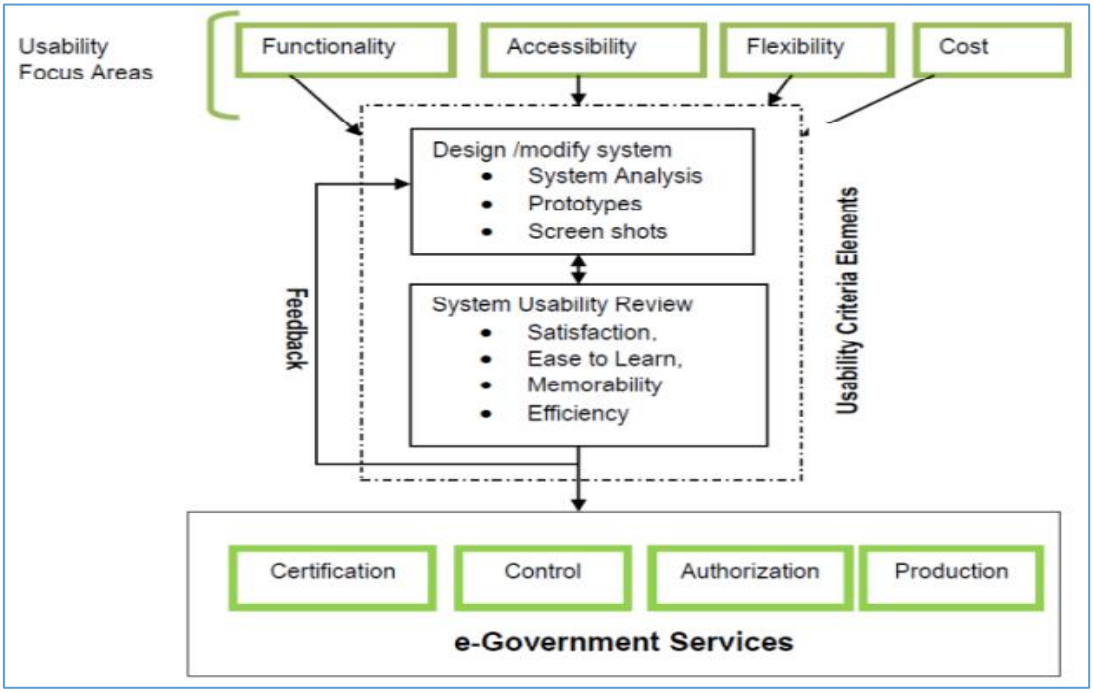



Figure 20: Framework for usability of e-Government services [30].

According to Mkude C., Wimmerthe M., these authors confirm that the examination of e-government implementation strategies in developed and developing nations yields results that are in line with the literature and with global surveys: Developing nations still lag behind industrialized nations by a wide margin. All contacted citizens said that e-government sustainability was an important aspect that needed to be taken into consideration in every e-government initiative. All respondents who took part in the process indicated a total of 24 sustainability-related criteria. The findings are shown in below table 8 [31].

| **Sustainability factors** | **In place in (country)** | **Recommended in (country)** |
|---|---|---|
| Commitment of government for implementation of e-government | UK, DK, SA, DE, GE | |
| Stable and continuous financing of various initiatives for the advancement of electronic services in the framework of electronic government | UK, DE, CH, MT, GE | TN, EG, MW, LB |
| Constant supervision and maintenance of advanced solutions | CH, SA, NL, DE | TN, LB |
| To guarantee quality and interoperability, many standards with more recent developments are used | DE, UK, MT, RU | MW, NG, TN,GE, MX, EG, LB |
| Promoting transparency for the projects throughout implementation and evaluation | NL, AT | |

| | | |
|---|---|---|
| Continued support for the legal framework around the implementation of e-government policies | AT, DE, RU, GE | MX, NG, LB |
| Implementing e-government at the national level and coordinating it through a centralized agency that handles the practical execution of tangible technical solutions | SA, GE | TN |
| The symbiotic relationship between e-government plans and national development objectives and guidelines in fields including health, education, justice, prosecution, and other critical national areas | DK, MX | UK |
| Projects should be evaluated annually by relevant parties for their outcomes, priorities, and sustainability | AT | |
| Robust business cases should be used | DK | |
| Use of precise and strict criteria for contracts, as well as for project development and execution processes | DK | |
| Development and advancement of recurring solutions | SE | |
| Ensures coordination of relationships and linkages among various, particularly important projects' strategies and plans | NL | |
| Track usage of e-services and feedback of users | MT | |
| Regardless of the political leadership shift, continuous commitment and support is provided by political level | DK | NL |

| | | |
|---|---|---|
| Gather, apply, and spread information about e-government deployment within the public sector | SE | |
| Ensure sufficient ICT infrastructure | RU | |
| Enlist the assistance of professionals that have a long-term viewpoint on e-government solutions. | | SE, EG, LB |
| Presence of a centralized evaluation framework | | NG, MW, EG |
| Ensure sufficient ICT capacity in public sectors | | GE, EG, MW |
| Ensure cooperation and coordination between ministries | | DE, EG |
| Use accountability procedures while outsourcing tasks to private sectors | | UK |
| Combining and using the expertise and knowledge of groups of researchers together with public sector officials who practice this knowledge continuously in public institutions | | MT |
| Citizens' request that the government provide the most qualitative and advanced electronic services | | MX |

Table 8: E-government sustainability factors named by respondents [31].

According to Kettani, D., there are eight characteristics that through them it is possible to outline the qualities or elements of good governance. These characteristics or components are as shown in Figure 21.

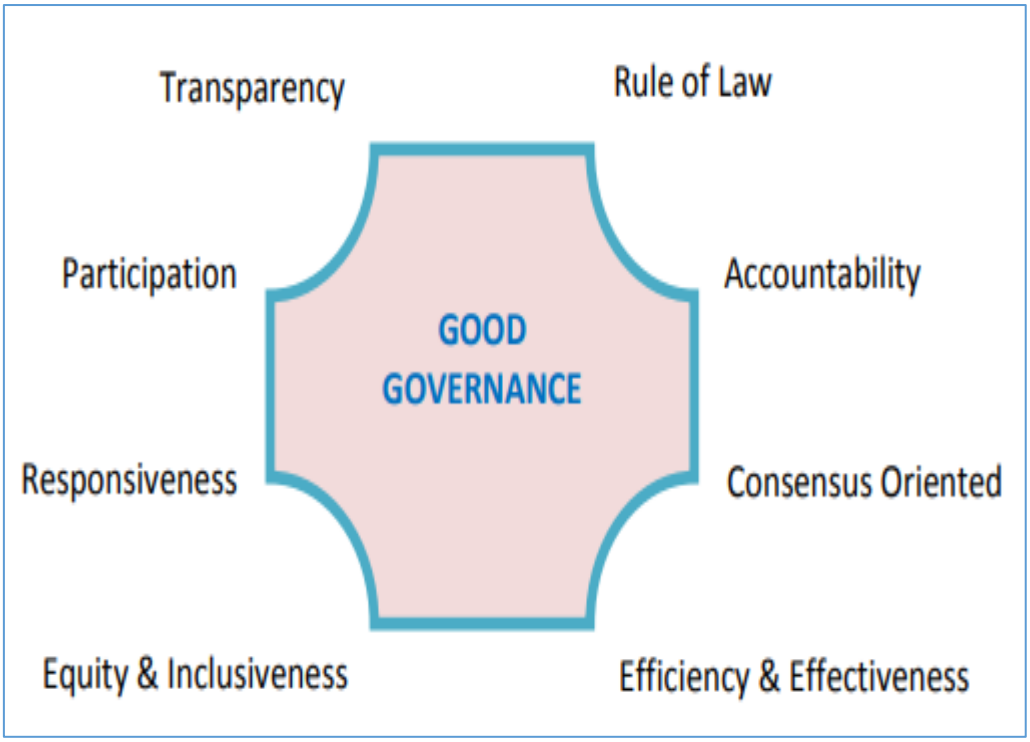



Figure 21: Characteristics or components of good governance [32].

According to Mohammad Abdul Salam, in Figure 22 it is presented that the conceptual analysis, theoretical foundation, and empirical research served as the basis for the development of the analytical framework. The model uses variables for measuring citizen satisfaction about delivery of e-services to citizens that are developed under the perspective of good governance. The analytical framework shows how qualities of customer satisfaction and contingent elements of good governance were used to evaluate how well the citizens were served throughout using the e-services. [33].

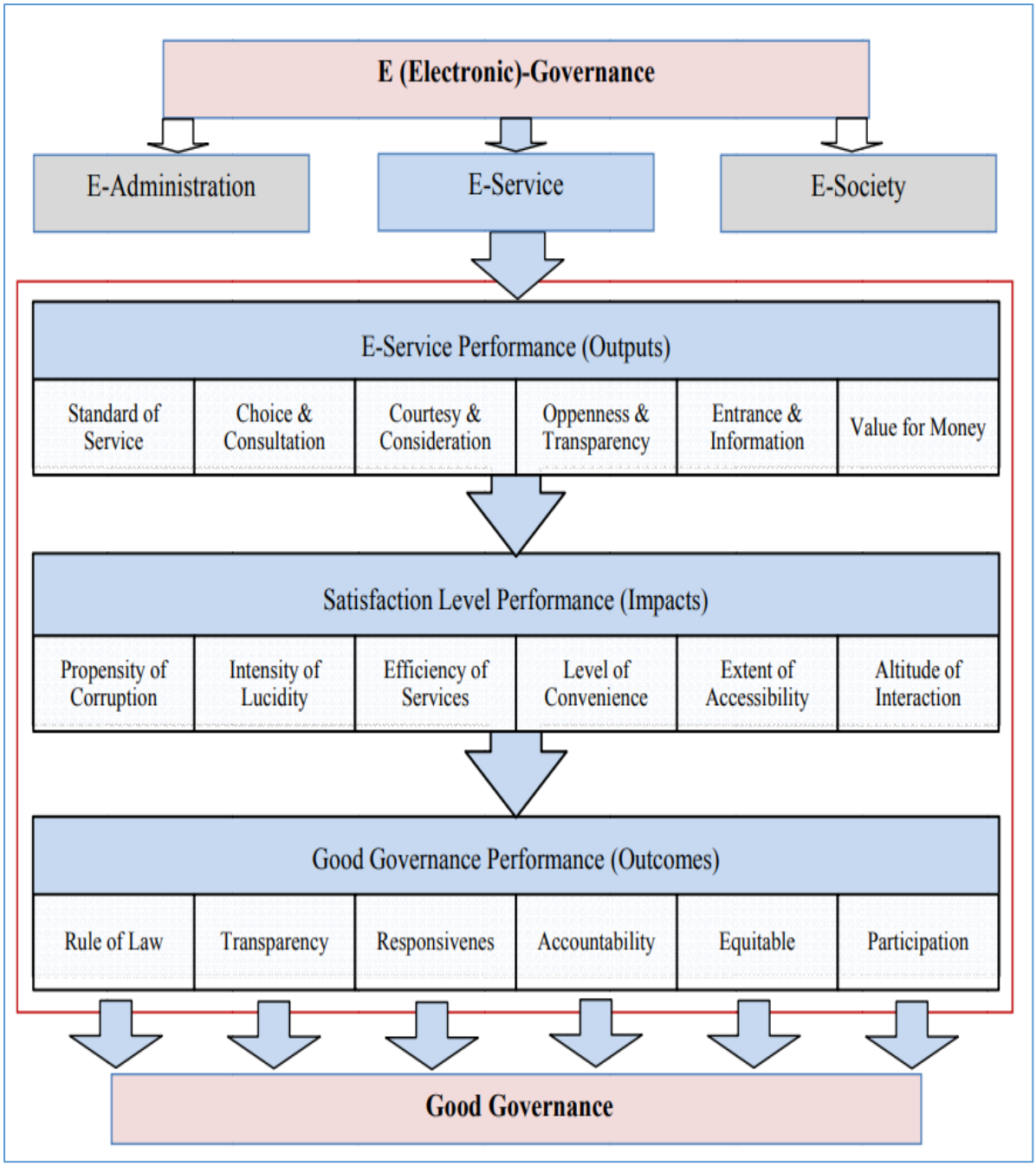



Figure 22: Analytical framework of e-governance for good governance [33].

According to Syed Faizan Hussain Zaidi, Mazen K. Qteishat, the authors confirm that based on multiple assessments, there is always a need to comprehend and evaluate consumers' impressions of e-government, and it is always crucial to grasp what makes a high quality electronic service. Governments are viewed as service providers, thus providing high-quality services will guarantee that people are happy and satisfied with government work. It is also obvious that more study is necessary to fully comprehend and interpret the results of the

recent studies on people' perceptions of the components of e-service quality. Therefore, the necessary parameters for the creation of an e-governance service quality assessment framework should be defined in order to determine the quality of the electronic service (e-GSQA). The proposed framework specified as e-GSQA for evaluating delivery of e-services is shown in Figure 23 [34].

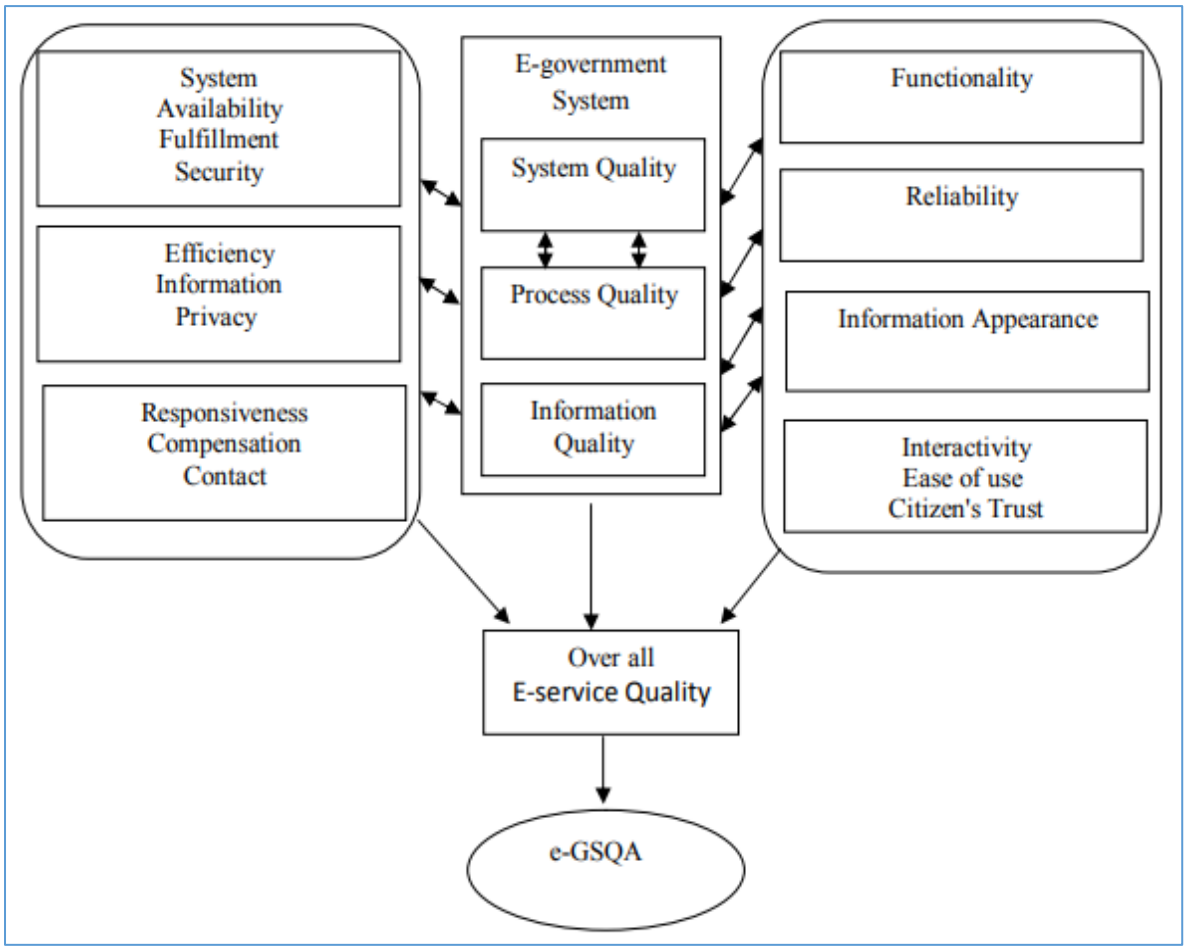



Figure 23: e-GSQA framework for assessing e-service delivery [34].

A specific and comprehensive model is proposed by Mahlangu G. and Ruhode E. Through this model author's try to describe notified service gaps during assessment of e-Government. The proposed model suggests that there are three different ways that e-government service gaps might be assessed:

- Functionality
- Delivery

- Service gaps.

These dimensions are shown in Figure 24 together also with their measurement components [35].

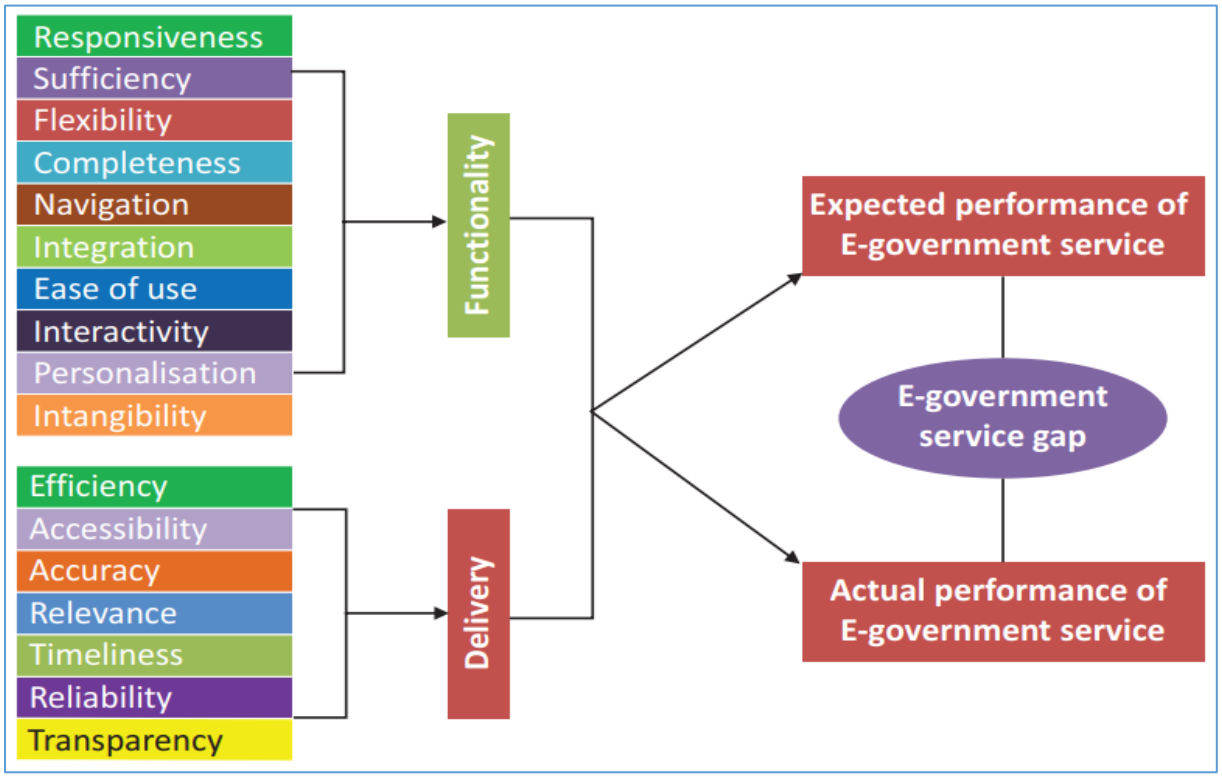



Figure 24: E-government service gap assessment model [35].

## 2.9 Summary

This chapter has presented and discussed related work on data quality dimensions and indicators, main problems and ways to improve data quality. Also in this chapter, we presented frameworks and methodologies that are proposed by different authors for managing, assessing and improving data quality. We presented main algorithms for interconnections of multiple large datasets. We also presented the importance of offering qualitative e-services by government institutions with the main condition of having qualitative data as a prerequisite. By considering these approaches, we have identified the requirements that should be fulfilled in this thesis.

# 3
# Data Quality Frameworks

## 3.1 Introduction

Certain important data and information currently support some of the most important business choices in a very substantial way. Having high quality data helps an organization to establish its best strategy and helps in making most adequate decisions. Because data may not present a clear and accurate picture of the situation, low data quality levels can have significant repercussions on a corporation, including poor decision-making and also missed business opportunities [55, 56].

Correcting the data stored in the database of different and crucial electronic registers in the middle of their processing process may increase the cost. The increasing volume of data available with unknown levels of quality of this data continues to add challenges to optimally analyzing the use of data that is important to the organization. Additionally, data warehouses have risen in size and complexity as a result of expansion of different organization's recently and ongoing in the number of data sources [57, 58]. The required and adequate decision-making process includes processes for obtaining, storing, distributing, applying, and truncating of the data [5]. A variety of techniques and procedures are implemented in the process of assessing data quality and the process of enhancing data quality by using appropriate methodologies and methods. This is because it is crucial to get and provide the best quality of data.

Therefore, it is crucial that all key data users, data consumers and other stakeholders involved in data input, data processing, and data analysis to participate these processes of evaluating data quality at all stages.

This chapter provides a summary of data quality frameworks that are generally applicable by comparing their main elements, including the process of data quality definition, the

process of data quality assessment and finally the process of data quality improvement. In addition, we attempt to provide advice on how to identify a suitable framework for evaluating and enhancing DQ. Section 3.2 presents and discusses structure and type of the data. Data quality frameworks are presented in Section 3.3. Section 3.4 treats data quality measurement and assessment. Section 3.5 treats the process of data quality improvement. Section 3.6 presents choosing appropriate data quality frameworks for specific datasets. Section 3.7 summarizes this chapter.

## 3.2 Data Definition

Data definition tries to describe the manner in which the data are determined in various data sets and their quality during the development of a methodology are thought to be the first step, which can also be described as a prerequisite and which in and of itself must include the context of the data, the nature of the data, and the data type.

### 3.2.1 Structure and Types of Data

There are two primary definitions for the nature of data and data type that we must take into consideration while attempting to describe the structure and type of the data that we want to assess and improve [57, 60].

The first definition classifies certain data into three categories: raw data item, component data item and information product.

In table 9 it is shown the description for each classification that is specified in first definition [57].

| **Data type** | **Description** |
|---|---|
| Raw Data Items | In this category are data that have not changed since they were processed or created and stored in storage |
| Information Products | This categorization is the outcome of manufacturing |

| | |
|---|---|
| | operation done with data |
| Component Data Items | This category is created each time the related information product is needed, and it is also temporarily held until the finished product is produced |

Table 9: Data types.

According to the second definition, data are classified as structured, semi-structured and unstructured [57, 61-63]. Table 10 shows this [64].

| **Data structure** | **Description** | **Example** |
|---|---|---|
| Structured Data | It refers to the groupings or generalizations of objects having basic or elementary qualities that are defined inside a domain | Tables with relational data |
| Semi-Structured Data | Through this classification, data is described with a structure that has a degree of flexibility | XML files |
| Unstructured Data | According to this classification, data structure is a generic sequence of symbols, typically coded in natural language | In e-mails is the body of text |

Table 10: Data structure.

Structured data is the focus of the majority of approaches for evaluating and improving the quality of data, while semi-structured data that has some flexibility is also taken into account. Except for HDQM, which is a rare approach that specifically takes into account structured, semi-structured, and unstructured data, many frameworks that can analyze structured and semi-structured data find it difficult to deal with unstructured data.

### 3.2.2 Poor or Dirty Data

Data quality issues can have a negative influence on the success of many organizations and institutions and have a range of ramifications, including social and economic ones [65]. Below are described the primary effects that can arise from both a conceptual and organizational standpoint [62]:

- **Operational**: Dissatisfaction of employees and customers as well as rising operating costs
- **Tactical**: It has huge effect when dealing with making decisions, a lack of trust inside the organization
- **Strategic**: It assesses how increased difficulty affects when we try to define and execute strategies.

## 3.3 Data Quality Frameworks

Through this dissertation, we hope to provide an overview of the different and complete data quality frameworks and offer advice on how to select the best one for a particular evaluation and improvement of data quality.
In general, each framework should have components for each of the main processes that have been defined:

- DQ definition
- DQ assessment
- DQ improvement.

First and foremost, the frameworks used should be broadly adaptable to the data context, information system, and all types of business.

A data quality methodology typically consists of three steps or phases that must be done in order for the activities to be completed [57]:

- *State reconstruction*, which has as its main responsibility gathering data on organizational procedures and services, relevant information, data quality issues and corresponding expenses.
- *Assessment/measurement,* which uses quality indicators to evaluate data quality of collections.
- *Improvement*, which is about choosing the actions, plans, and tactics to achieve new data quality objectives.

The main viewpoints that can evaluate and compare most important data quality techniques are listed below [57]:

- the *phases and steps* which make up the methodology
- the *strategies and techniques* that are used in the process for evaluating and boosting data quality levels
- the *dimensions and metrics* which are selected in the approach to evaluate the degrees of data quality
- the *types of costs* that are related to problems with data quality, such as the price of bad data and the price of actions for evaluation and improvement
- the *types of data* that the technique takes into account
- the *types of information systems* that make use of, alter, and manage the data used in the approach
- the *organizations* part of the procedures that produce or update the data used in the technique, including its structure and standards
- the *processes* that generate or update data with the intention of generating services that customers need and are taken into account by this approach
- the *services* that are produced by the procedures taken into account by the methodology.

Table 11 displays various frameworks in accordance with the procedures for the definition, assessment, and enhancement of data quality [64].

| **Acronym** | **Framework** |
|---|---|
| AIMQ | A Methodology for Information Quality Assessment |
| CDQ | Comprehensive Methodology for Data Quality Management |
| COLDQ | Cost-effect of Low Data Quality |
| DQA | Data Quality Assessment |
| DQAF | Data Quality Assessment Framework |
| DQPA | A Data Quality Practical Approach |
| HDQM | A Data Quality Methodology for Heterogeneous Data |
| HIQM | Hybrid Information Quality Management |
| OODA DQ | The Observe-Orient-Decide-Act Methodology for Data Quality |
| TBDQ | Task-Based Data Quality Method |
| TDQM | Total Data Quality Management |
| TIQM | Total Information Quality Management |

Table 11: Description of frameworks and components [64].

Described frameworks in the table above are generally implemented, there are also other frameworks that are focused only on a particular objective, such as CIHI (Healthcare Data),

AMEQ (Manufacturing Product Data), IQM (Web Data), MAMD (ISO Standards), etc., which, due to their distinct purpose, are not addressed in this dissertation.

## 3.4 Data Quality Measurement and Assessment

The process of evaluating data quality may be characterized as the measurement of the effects of various business activities with focus on the data [67]. The management of data quality requires defining in detail this process which involves a variety of procedures and personnel in the company or institution. Additionally, we should take into account the source data, the process that created the data item, and data aggregation when performing the data quality assessment. [52].

### 3.4.1 Data Quality Measurement Ways

There are two methods that data quality can be measured:

- **Subjectively**, using the data consumer's evaluation and judgment of the quality of the dimensions
- **Objectively**, for instance, by utilizing various calculations that might indicate the level of data quality, such as calculating the percentage of stated restrictions or the number of incorrect judgments [68].

The majority of metrics used to assess data quality typically range from 0 to 1, with 0 denoting an inaccurate result and 1 denoting a proper value.

The following function calculates these attributes or dimensions, such as completeness, accuracy, and consistency:

$$F=1-(Fi/Ft) \quad (1)$$

Where Fi is the number of wrong values, Ft is the total number of values for the relevant dimension, and F is the metric for that dimension.

If we want to calculate completeness dimension, we can do that by using the following formula:

*Completeness= 1- (No. of incomplete val. / total no. of val.)* (2)

In order to determine how accurate and error-free a set of data is, we can assess its accuracy by the following formula:

*Accuracy= 1 - (No. of val. in error / total no. of val.)* (3)

Since some aspects of data quality cannot be evaluated using objective metrics, they should be evaluated using subjective metrics [69].

Table 12 displays a summary of measurement types used in every framework, where the majority of frameworks include objective data quality measurement [64].

| Framework | Main Components |
|---|---|
| AIMQ | Questionnaire for survey |
| CDQ | Definition of data quality and accuracy metrics for user interviews |
| COLDQ | Consumer surveys and various definitions of data quality metrics |
| DQA | Definition of quantitative metrics and expectations of stakeholders |
| DQPA | Definition of the primary data source with derived data quality metrics |
| DQAF | Metrics for different forms of measurement's data quality |
| HDQM | Defined metrics for accuracy and currency of data quality |
| HIQM | Objective evaluation when a measuring algorithm is suggested |
| OODA DQ | It is not specified |

| | |
|---|---|
| TBDQ | Simple survey questions and ratio |
| TDQM | Business guidelines and metrics for data quality |
| TIQM | The definition of data quality measurements and user expectations |

Table 12: Measurement types.

The DQA framework presents the functional forms for the evaluation of objective data quality criteria presented in Table 13 [64].

| **Functional Form** | **Description** | **Dimensions measured** |
|---|---|---|
| Simple Ratio | Intended results to ratio of overall results | Error free, completeness, consistency, brief presentation, relevance |
| Min or Max Operation | A normalized individual data quality indicator's minimum or maximum value | Credibility and the right amount of information |
| Weighted Average | The process of allocating weighting factors to signify the variables' significance in the assessment of a dimension | Credibility and the right amount of information |

Table 13: The DQA Framework's functional forms.

### 3.4.2 Assessment Steps for DQ

When we want to assess data quality in datasets, in this phase we have to follow steps listed below [57]:

- *Data analysis*, which aims to fully comprehend rules, processes and data
- *Data quality requirements analysis*, which, via the use of data, administrators, and users, detects issues with data quality and establishes new data quality goals
- *Process for Identification the most critical areas*, which in quantitative way evaluates the most significant databases and flows of data
- *Modeling of Process*, which generates or updates data using the designated model
- *Quality measurement*, which as a result sets specific metrics of the quality dimensions that are impacted by the DQ requirements analysis.

It has been shown that some frameworks, including TDQM and DQA, attempt to apply specific measurements of data quality even if they do not offer clear instructions for the assessment process.

The evaluation phase is structured differently depending on the framework chosen, however there are many common elements between the different frameworks in terms of types of measurements.

## 3.5 Improvement Process of DQ

After the process of determining the data's quality, there are different pertinent strategies and practical tools that should be considered to apply with the aim of enhancing the quality of these data, where there are several phases that should be considered to implement.
Below are listed two main categories of strategies that frameworks apply most of the times:

- Strategy of data-driven
- Strategy of process-driven.

***Strategy of data driven*** improves DQ directly by changing the value of the data. The improvement approaches used in this strategy include the acquisition of initial data, standardization (or normalization), record linking, data and schema integration, source reliability, error localization and rectification, and cost optimization [57].

***Strategy of process driven*** specifies that in order to improve data quality, procedures that produce or edit data should be redesigned all the time based on the needs. Process control and Process redesign are the two key improvement strategies that define this strategy for improvement of DQ [57].

The actions listed below should be taken during the period or phase of improving the data quality [57]:

- *Cost evaluation step*, which step is required to assess the direct and indirect costs of data quality improvement
- *Assigning process duties*, which defines the process owners and notifies their duties for data management and production operations
- *Assigning data duties*, which specifies the data owners and notifies their duties for data management and production operations
- *Process of identifying the reasons of possible mistakes*, which examines the root causes of quality issues
- *Selection of approaches and strategies,* which indicates the appropriate procedures and the required data enhancement strategies
- *Design of data improvement solutions*, which chooses the most effective tactics, methods, and equipment with the intention of enhancing data quality
- *Control of the Process*, which provide checkpoints in the data generation operations to monitor quality
- *Redesign of the process*, which specifies steps for process improvement that might lead to comparable improvements of data quality

- *Improvement management*, which establishes new guidelines and rules for improvement of data quality inside organizations
- *Improvement monitoring*, which, via a series of monitoring actions, offers feedback on the outcomes of the improvement process.

Several frameworks, including the TDQM and DQA frameworks, do root cause analyses of data quality as a first step to improving data quality.

## 3.6 Choosing effective Framework for Assessment and Improvement of DQ

When evaluating and enhancing data quality, various data quality frameworks and methods are used to do this process. In the section below, we will provide guidance that is meant to aid data science researchers in selecting the best framework for data quality in a certain circumstance.

Table 14 lists a few preliminary actions that must be taken before using the guide to make any choice or decision:

1. Several frameworks have been established for specific proposals if user expectations and needs are for a particular field of data quality. Frameworks for this specific purpose are provided in Table 14 together with a very specific application [64].

| **Framework** | **Regarding a certain objective** |
|---|---|
| AMEQ | Focuses on Product Manufact. Data |
| CIHI | Handles information created by electronic systems in healthcare |
| DaQuinCIS | provides treatment for data generated by Cooperate Information Systems (CIS) |

| | |
|---|---|
| DWQ | Addresses the warehouse's quality of data |
| IQM | Focuses on the web's data quality |
| ISTAT | Handles Census data |
| MAMD | Treats Standards for Int. Organization |
| MMPRO | Treats Standards for Int. Organization |
| ORME-DQ | Management of Operational Risk |
| PDQM | Focuses on the web's data quality |
| ProDQM | Addresses the warehouse's quality of data |
| QAFD | Handles data quality in financial data field |
| UDQA | Analyze data quality from a utility-driven approach |

Table 14: Summary of specific data quality frameworks.

2. Here are some potential possibilities for frameworks if the user's needs are for general frameworks that provide extensive help for the defining, analyzing, and improving processes of treating quality of data.
3. All of the frameworks provided here can be used in many specific contexts depending on the context of the data being processed (aside from the frameworks in Table 13 that are for a specific application).

The number of data quality dimensions that are used by each framework with widespread application is shown in Table 15 [64].

| No. | Framework | No. of Implemented Dimensions |
|---|---|---|

| | | |
|---|---|---|
| 1. | AIMQ | 14 |
| 2. | CDQ | 15 |
| 3. | COLDQ | 34 |
| 4. | DQA | 16 |
| 5. | DQAF | 5 |
| 6. | DQPA | 7 |
| 7. | HDQM | 2 |
| 8. | HIQM | 4 |
| 9. | OODADQ | 2 |
| 10. | TBDQ | 4 |
| 11. | TDQM | 15 |
| 12. | TIQM | 16 |

Table 15: Data quality's dimensions number used in frameworks [64].

The analysis of data and quality measurement are the processes that these frameworks utilize the most frequently, according to Table 16, which compares how each one is used during the evaluation phase.

| No. | Framework | Data Analyses | DQ Req. Analyses | Identif. of Critical Areas | Process Model. | Measurem. of Quality |
|---|---|---|---|---|---|---|
| 1. | AIMQ | + | - | + | - | + |
| 2. | CDQ | + | + | + | + | + |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3. | COLDQ | + | + | + | + | + |
| 4. | DQA | + | - | + | - | + |
| 5. | DQAF | + | + | + | - | + |
| 6. | DQPA | + | + | + | - | + |
| 7. | HDQM | + | - | + | + | + |
| 8. | HIQM | + | - | + | + | + |
| 9. | OODADQ | + | - | - | + | + |
| 10. | TBDQ | + | + | - | - | + |
| 11. | TDQM | + | - | + | + | + |
| 12. | TIQM | + | + | + | + | + |

Table 16: Steps and frameworks for assessing the quality of data.

The comparison of framework improvement stages is shown in Table 17, where the method for detecting error sources is the improvement step that is most frequently implemented.

| No. | Framework | Eval. of Costs | Selection of Strateg. and Techni. | Identif. of reasons of possible mistakes | *Redesign of the process* | Improve. Monitoring |
|---|---|---|---|---|---|---|
| 1. | AIMQ | - | + | + | + | + |
| 2. | CDQ | + | - | + | - | - |
| 3. | COLDQ | + | + | + | - | + |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4. | DQA | - | - | + | - | - |
| 5. | DQAF | + | - | + | - | - |
| 6. | DQPA | + | - | + | + | - |
| 7. | HDQM | + | + | + | + | + |
| 8. | HIQM | - | - | - | - | + |
| 9. | OODADQ | - | + | - | + | - |
| 10. | TBDQ | + | + | + | + | + |
| 11. | TDQM | + | + | + | + | + |
| 12. | TIQM | + | + | + | + | + |

Table 17: Steps and frameworks for improvement of the quality of data.

In the following Table 18, it is provided the guide framework for choosing the best suited framework.

| **Question** | **Answer** | **Applic. Frameworks** |
|---|---|---|
| How are structured the data that is being processed? | Struct. data | 1,2,3,4,5,6,7,8,9,10,11,12 |
| | Semi-struct. data | 1,2,3,7,8,9,11,12 |
| | Unstruct. data | 1,3,7,9,11,12 |
| What kind of | Objective metrics | 2,3,4,5,6,7,10,11,12 |

| | | |
|---|---|---|
| metrics are to be used with the processed data? | Subject. assessments | 1,2,4,10,12 |
| Is it supported the process of identifying the dimensions? | Yes | 2,4,8,10 |
| | No | 1,3,5,6,7,9,11,12 |
| What are the most relevant dimensions? | Metrics n Table 15 | Supporting frameworks for certain dimensions |
| What steps are used during the evaluation process of data quality? | Data Analyses | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 |
| | DQ Req. Analyses | 2, 3, 5, 6, 10, 12 |
| | Identif. of Crucial Areas | 1, 2, 3, 4, 5, 6, 7, 8, 11, 12 |
| | Process Modeling | 2, 3, 7, 8, 9, 11, 12 |
| | Measurement of Quality | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 |
| What steps are used during the improvement process of data quality? | Cost evaluation step | 2, 3, 5, 6, 7, 10, 11, 12 |
| | Selection of approaches and strategies | 1, 3, 7, 9, 10, 11, 12 |
| | Process of identifying the reasons of possible mistakes | 1, 2, 3, 4, 5, 6, 7, 10, 11, 12 |
| | Redesign of the process | 1, 6, 7, 9, 10, 11, 12 |

| | | |
|---|---|---|
| | Improvem. Monitoring | 1, 3, 7, 8, 10, 11, 12 |

Table 18: Process of choosing a suitable framework for handling data quality.

## 3.7 Summary

In this chapter, we presented an overview of data quality frameworks by surveying and comparing types and structure of the data. Row, component, and information product data types are the data types that are displayed. Structured, semi structured, and unstructured data types have all been treated in this chapter. We also reviewed and compared the many data quality metrics used for different frameworks, where certain metrics are found in many different frameworks while others are exclusive to only one. During the assessment and improvement of DQ, we came to conclusion after looking at all the variables that correctness, completeness, and timeliness appear to be the most crucial data quality characteristics.

Many approaches and techniques are discussed in this chapter, and it is made clear that each of these frameworks follows a set process described in specific steps for defining, evaluating, and improving data quality.

Through this chapter hypothesis 3 is confirmed that the proposed model of data quality improvement positively influences the quality of data output.

In order to increase the quality of datasets and electronic services, the offered overview highlights the significance of assessing and improving data quality by selecting the most suitable dimensions and frameworks.

# 4

# Data Cleansing Challenges and Techniques

## 4.1 Introduction

Since data is a very important asset for various organizations, government institutions, companies, universities, etc., good data quality helps decision makers for making the right decisions in order for citizens to receive better e-services and to enable the automation of many processes. Having data in the best shape, the data and data groups that have many anomalies must be treated through the data cleansing process as a very sensitive process. Anomaly as a property of data values affects the incorrect representation of the state of the processes and resulting from erroneous measurements, wrong data inputs, eventual omissions occurring during data processing, etc. [83].

In order to present the data as well as possible and integrate them with other systems, these data must go through the process of cleansing them from possible errors. This includes many auditing processes to find errors and also processes to fix those errors. Different organizations and companies have different approaches and methods for cleansing of data.

According to Rahm, E., data quality problems can be divided into single-source problems that include record, record type, attribute and multiple-source problems that include naming conflicts, schema-level problems and references of the same entity [84].

There is a number of anomalies, but below we are listing some of the most important anomalies that should be taken into account:

- Order dependencies violation
- Negative numbers reported
- Delayed-reported issue on weekend/holiday
- Abnormal data point or data period.

We can understand the real state of the current data through the process of data quality assessment and data visualization, specifically through the inclusion of dimensions such as, accuracy, timeliness, completeness, relevancy, etc.

We can often come across two terms, data quality and data reporting quality, they are not completely the same. Although accuracy is an essential element in the quality of data, the emphasis is not on this element when measuring the quality of data reporting, but the main aspect is the presence or absence of a piece of information as well as the format in which it is reported [39].

To better understand the data, visualization is the most suitable form, where using appropriate statistical graphics, even complex ideas can be explained clearly and accurately.

The case study on the procedure of gathering, integrating, comparing, cleaning, evaluating, and visualizing datasets linked to the global disease known as COVID-19 is presented in this chapter.

The chapter is organized as follows starting with introduction in section 4.1. In section 4.2 we first overview data cleansing anomalies encountered in datasets. In section 4.3, we will analyze official COVID-19 datasets like World Health Organization (WHO), John Hopkins University (JHU), European Centre for Disease Prevention and Control dataset (ECDC), Republic of North Macedonia Dataset (RNM) and National Institute of Public Health in Kosovo (NIPHK). In section 4.4, we will examine techniques and illustrate how to gather, evaluate, improve, and visualize the data for various datasets. In section 4.5, we will present experimental results collected from the assessed datasets. Section 4.6 summarizes this chapter.

## 4.2 Abnormal Data Detection

Data inconsistencies or anomalies that were discovered when examining datasets are:

- Negative numbers
- Order dependencies violation
- Delayed-reported issue on weekend/holiday
- Abnormal data point or data period.

For the data to be reliable, these abnormalities must be removed and corrected.

A. NEGATIVE NUMBERS

In different types of datasets, such is COVID-19 dataset should not contain negative values because the statistics for verified cases or fatalities should be zero or any other positive number. For this reason, these datasets must be cleaned from negative numbers that may have been mistakenly reported.

B. ORDER DEPENDENCIES VIOLATION

In order to detect abnormal data and repair cumulative time series data, the concept of order dependency (OD), which is also widely used in relational databases, will be incorporated. In the case of the COVID-19 datasets, the following definition applies to OD for the cumulative time series: for any two time points, t1 and t2, if $t1 < t2$, then $Yt1 \leq Yt2$, where Y represents the cumulative infectious/death count [44].

C. WEEKEND/HOLIDAY DELAY-REPORTED ISSUE

In some datasets, it can be seen that there is a much smaller number of new daily cases reported during weekends and holidays, and this issue is called weekend/holiday delay-reported.

D. ABNORMAL DATA POINT OR DATA PERIOD

There are situations when on a certain day there can be an abrupt increase in the cumulative time series of new cases or deaths and this process is called the abnormal data point. This situation can be caused because of:

1. A significant number of tests' findings were published
2. The modification in reporting requirements, for instance, the start date at which some states began to declare probable cases.

There are occasions when a continuous anomalous phase may be seen, meaning that the growing pace is noticeably different from the prior and succeeding periods.

The typical workflow for data curation is presented in Figure 25 [44]. As can be understood from the figure below, the detection and repairing of the OD violation for raw data is performed firstly, then continuing with the next step, which is the check for delay-reported issues on weekends/holidays. The last step is checking for the existence of abnormal data points and periods.

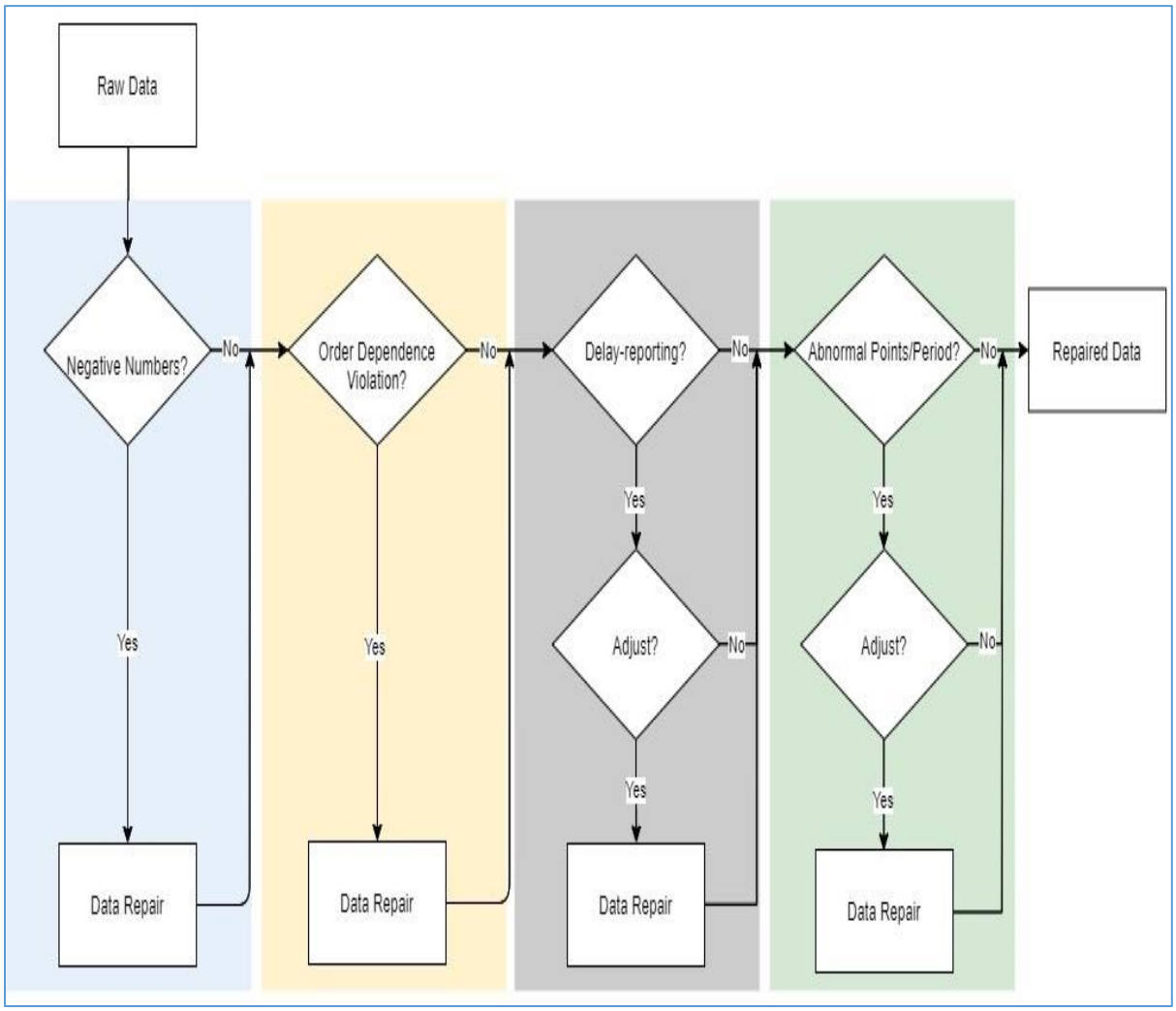



Figure 25: Data curation flowchart [44].

Due to the rapid global spread of the COVID-19 virus as well as delayed reporting, COVID-19 datasets in particular have data quality issues [45].

In our case study, the datasets used are of Covid-19 from the years 2020 and 2021 to apply data cleansing techniques with the aim of data quality assessment and improvement for a better prediction and to be more prepared with the problems they could face.

There are two types of data cleaning procedures:

1. Cleaning of data manually
2. Cleaning of data automatically.

Although manual cleaning has quite high accuracy, it is nevertheless problematic to implement because it requires a lot of commitment and takes a lot of time.

In order to analyze the data and visualize the comparative results of these datasets, the Power BI software was used.

## 4.3 Experimental Datasets

As we know, the Covid-19 virus has spread for a short time all over the world since the first case reported in Wuhan, so global cooperation and coordination has been imperative in order to deal with this virus.

In addition, the consistent evidence of data for COVID-19 is critically important in order to understand transmissibility and transmission routes, geographical spreading as well as the risk factors for infection [37].

Providing accurate and trustworthy data proof has several advantages for government organizations, health institutions, scholars, and other interested parties. The main benefit is for the healthcare system to assess the impact of the pandemic and to create appropriate policies and better planning [38].

### 4.3.1 Analysis of data from Crucial Datasets

World Health Organization (WHO), the European Centre for Disease Prevention and Control (ECDC), John Hopkins University (JHU), the Republic of North Macedonia, and the Republic of Kosovo were the sources of the data for this study.

### 4.3.2 World Health Organization (WHO) Dataset

The national authorities of the countries have continuously reported the cases of COVID-19 to the World Health Organization (WHO) on a daily basis, including raw data and metadata. For the purpose of more adequate representation and analysis of cases for different countries, the data were in Microsoft Excel and CSV format, although at the beginning these data were reported in pdf format [40]. At the beginning, it was difficult to read the data completely and automatically from the report in pdf format due to the different structure of pdf files and non-standardized tables. It was even necessary to manually interfere several times in the adjustment of the data. The fact that the datasets are now organized in Microsoft Excel and CSV formats has an impact on how easy it is to analyze, enhance, and visualize the data.

A sample World Health Organization dataset is shown in following table:

| Date reported | CountryCode | Country | WHO region | New cases | Cumulative cases | New deaths | Cumulative deaths |
|---|---|---|---|---|---|---|---|
| 3/11/2020 | AF | Afghanistan | EMRO | 3 | 11 | 0 | 0 |
| 8/27/2020 | AF | Afghanistan | EMRO | 55 | 38126 | 4 | 1401 |
| 3/9/2020 | AL | Albania | EURO | 6 | 6 | 0 | 0 |
| 8/24/2020 | AL | Albania | EURO | 152 | 8427 | 5 | 250 |
| 1/4/2020 | CN | China | WPRO | 1 | 1 | 0 | 0 |
| 8/24/2020 | CN | China | WPRO | 41 | 90182 | 1 | 4718 |
| 1/29/2020 | IT | Italy | EURO | 6 | 6 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 8/24/2020 | IT | Italy | EURO | 1209 | 259345 | 7 | 35437 |
| 3/13/2020 | XK | Kosovo | EURO | 6 | 6 | 0 | 0 |
| 8/24/2020 | XK | Kosovo | EURO | 111 | 12405 | 10 | 457 |
| 2/26/2020 | MK | North Macedonia | EURO | 5 | 5 | 0 | 0 |
| 8/24/2020 | MK | North Macedonia | EURO | 287 | 13595 | 7 | 564 |
| … | … | … | … | … | … | … | … |
| 3/20/2020 | ZW | Zimbabwe | AFRO | 1 | 1 | 0 | 0 |
| 8/24/2020 | ZW | Zimbabwe | AFRO | 37 | 5930 | 2 | 155 |

Table 19: A sample of World Health Organization dataset.

A World Health Organization dashboard is shown in Figure 26.

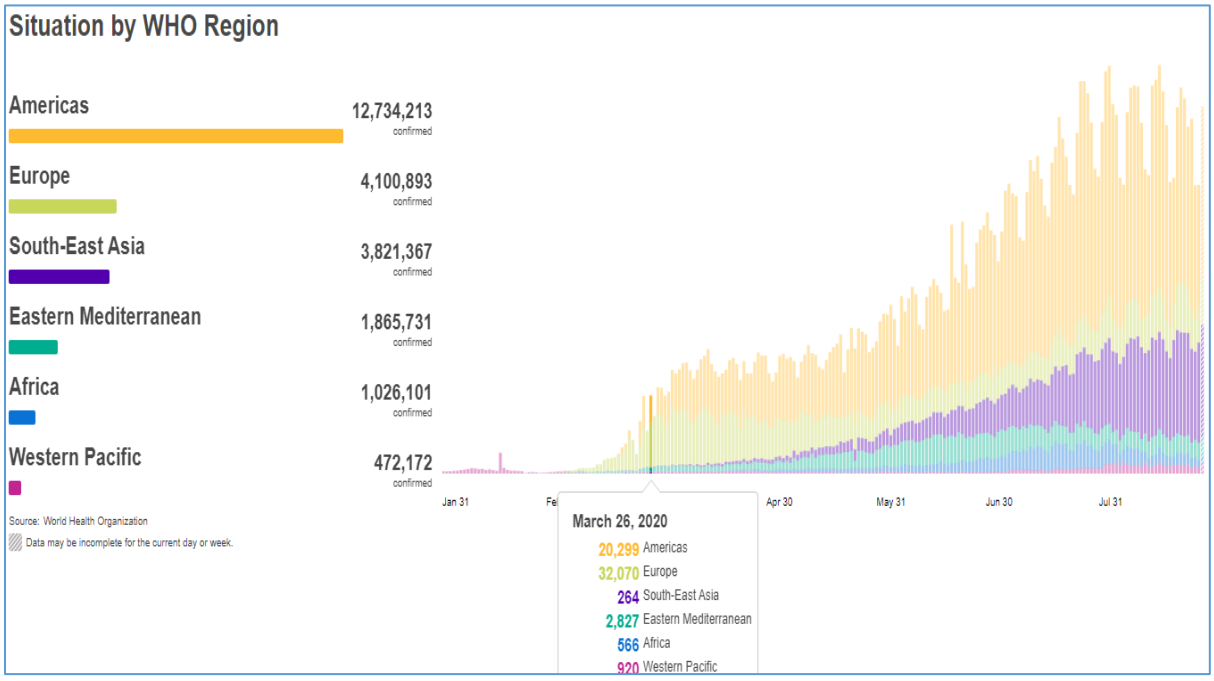



Figure 26: A dashboard from World Health Organization.

### 4.3.3 European Centre for Disease Prevention and Control (ECDC) Dataset

In ECDC website [41], there are datasets for COVID-19 in CSV, JSON and XML formats that contain the data fields as shown in the table below. These datasets contain the latest reported data on COVID-19 as they are updated every day, being continuously processed in order to ensure the accuracy and reliability of the data. This process helps not only in the European Union (EU), but also around the world, to monitor the dynamics of the COVID-19 pandemic.

The ECDC dataset data fields are in the following table:

| ISO code | Continent | Location | Date | total cases | new cases | total deaths | new deaths |
|---|---|---|---|---|---|---|---|
| BEL | Europe | Belgium | 2020-03-11 | 601 | 99 | 1 | 1 |
| BEL | Europe | Belgium | 2020-08-27 | 82936 | 0 | 9879 | 1 |
| BRA | South America | Brazil | 2020-03-18 | 291 | 57 | 1 | 1 |
| BRA | South America | Brazil | 2020-08-27 | 3717156 | 47161 | 117665 | 1085 |
| DEU | Europe | Germany | 2020-01-28 | 1 | 1 | 0 | 0 |
| DEU | Europe | Germany | 2020-08-27 | 237936 | 1507 | 9285 | 5 |
| OWID_KOS | Europe | Kosovo | 2020-03-14 | 2 | 2 | 0 | 0 |
| OWID_KOS | Europe | Kosovo | 2020-08-27 | 12683 | 0 | 488 | 0 |
| MKD | Europe | Macedonia | 2020-03-07 | 3 | 2 | 0 | 0 |
| MKD | Europe | Macedonia | 2020-08-27 | 13799 | 126 | 573 | 5 |
| … | … | … | … | … | … | … | … |
| ZWE | Africa | Zimbabwe | 2020-03-21 | 1 | 1 | 0 | 0 |
| ZWE | Africa | Zimbabwe | 2020-08-27 | 6251 | 55 | 179 | 13 |

Table 20: ECDC dataset.

### 4.3.4 Johns Hopkins University (JHU) Dataset

The Johns Hopkins Coronavirus Resource Center (CRC) continuously updates the data of the cases of COVID-19 as well as contains expert guidance. Through the collection and analysis

of data for COVID-19, such as new confirmed cases and deaths, they help policymakers and health authorities around the world in making decisions and implementing adequate policies for responding to the pandemic [36]. JHU files, which are in CSV format, offer a daily update of the pandemic's worldwide map.

The JHU dataset is shown in the following table:

| Province / State | Country/ Region | Lat | Long | 1/22/ 20 | 1/23/ 20 | … | 8/22/2 0 | 8/23/2 0 |
|---|---|---|---|---|---|---|---|---|
| | Afghanistan | 33.9391 1 | 67.709953 | 0 | 0 | … | 37953 | 37999 |
| | Albania | 41.1533 | 20.1683 | 0 | 0 | … | 8275 | 8427 |
| Queensla nd | Australia | -27.4698 | 153.0251 | 0 | 0 | … | 1105 | 1106 |
| Tasmani a | Australia | -42.8821 | 147.3272 | 0 | 0 | … | 230 | 230 |
| Victoria | Australia | -37.8136 | 144.9631 | 0 | 0 | … | 18231 | 18330 |
| | Belgium | 50.8333 | 4.469936 | 0 | 0 | … | 81468 | 81936 |
| | Japan | 36.2048 24 | 138.25292 4 | 2 | 2 | … | 61916 | 62658 |
| | Kosovo | 42.6026 36 | 20.902977 | 0 | 0 | … | 12168 | 12448 |
| | North Macedonia | 41.6086 | 21.7453 | 0 | 0 | … | 13458 | 13595 |

Table 21: JHU dataset.

A JHU dashboard for COVID-19 dataset is shown in Figure 27.

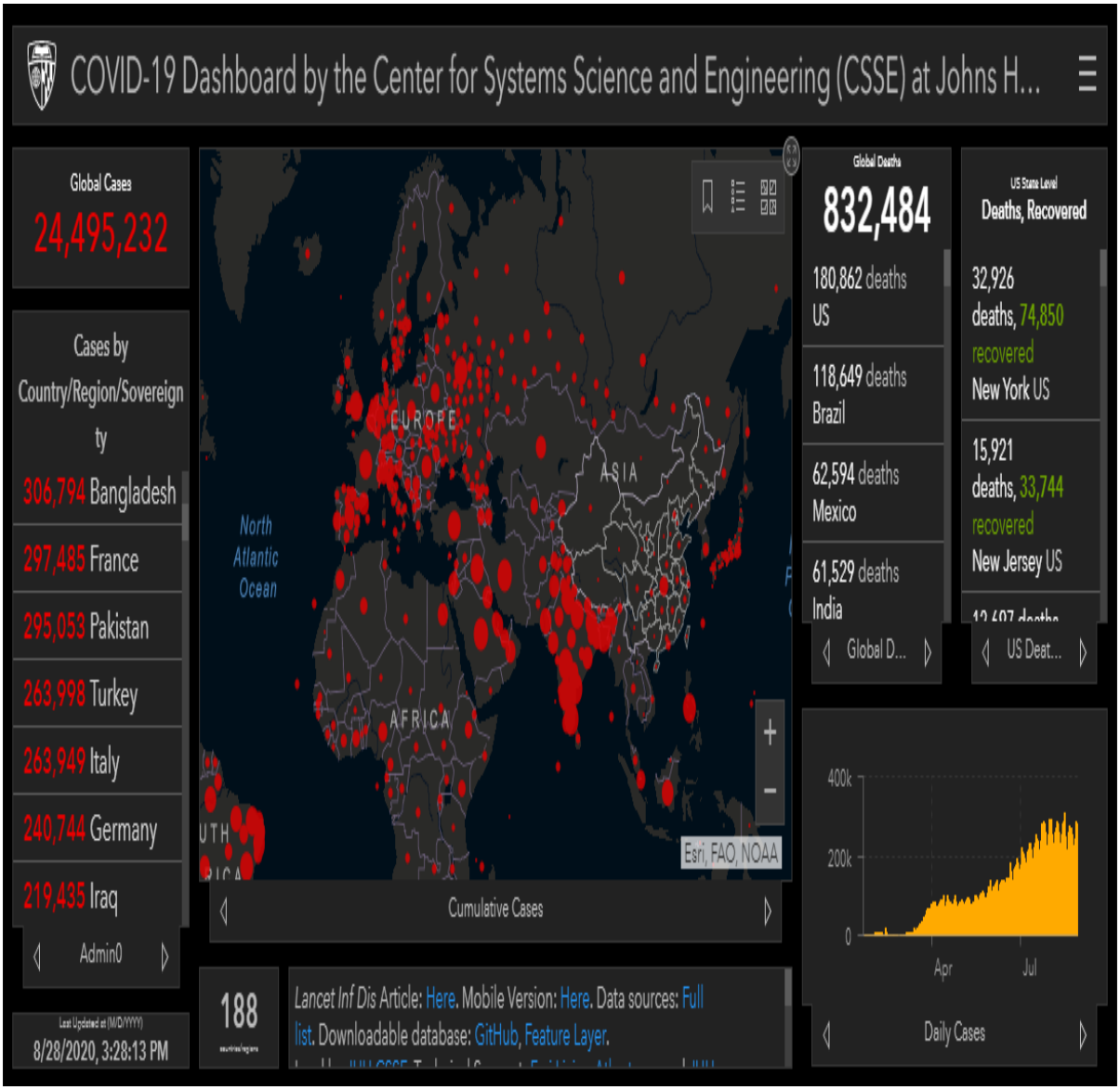



Figure 27: A dashboard from JHU for COVID-19 dataset.

### 4.3.5 Republic of Kosovo Dataset

The National Institute of Public Health of Kosovo (NIPHK) publishes data on COVID-19 on its official Facebook page [42]. The NIPHK dataset was created manually in order to be utilized for further research because the data is released as an image and not in a CSV, XML, or other appropriate format. The published data are the number of new cases, the number of

deaths, and the number of healed due to COVID-19, the cumulative values of new cases, deaths and healed.

The NIPHK data details are in the following table:

| COVID-19 cases by municipalities | Locality | Number of cases |
|---|---|---|
| Deçan, total: 2, by localities | Deçan | 1 |
| | Strellc i Ulet | 1 |
| Gjakovë, total: 19, by localities | Gjakovë | 14 |
| | Rogovë | 1 |
| | Skivjan | 3 |
| | Stubëll | 1 |
| Prishtinë, total: 23, by localities | Prishtinë | 20 |
| | Hajvali | 2 |
| | Matiçan | 1 |
| Prizren, total: 6, by localities | Prizren | 6 |
| …… | …. | …. |
| In total | 25 municipalities | 141 |

Table 22: Sample dataset from NIPHK.

### 4.3.6 Republic of North Macedonia Dataset

The Government of the Republic of North Macedonia (GRNM) publishes the results of cases with COVID-19 on its web portal [43], in tabular form and can be exported in formats such as CSV for analysis and comparison as needed.

The GRNM dataset details are in the following table:

| Date | Sick | Healed | Deaths in tot. |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 5-Mar-20 | 1 | 0 | 0 |
| 6-Mar-20 | 2 | 0 | 0 |
| … | … | … | … |
| 1-Apr-20 | 354 | 17 | 11 |
| 2-Apr-20 | 384 | 17 | 11 |
| … | … | … | … |
| 1-Aug-20 | 10900 | 6698 | 493 |
| 3-Aug-20 | 10066 | 6883 | 497 |
| … | … | … | … |
| 22-Aug-20 | 12592 | 10405 | 567 |

Table 23: Sample dataset from GRNM.

## 4.4 Methodology

In order to analyze and compare these datasets, initially these datasets had to be arranged in a unique appropriate form, because the data for new cases and fatalities in the JHU dataset are given in columns as opposed to rows in the other given datasets. Alpha-2 has been introduced as a new field to all rows of these datasets by fusing it with the Date variable in order to integrate the three datasets into a single dataset with a unique primary key for each nation.

A Joint Datasets Table is shown in following table:

| Attribute | Description | Additional information |
|---|---|---|
| Row_ID | ID of row | Unique ID |
| Date | Reporting date for Cases and Deaths | Date format in dd.mm.yyyy |
| Continent | Continent of the state | |
| Area | WHO Region | There are six world regions by World Health Organization, for the purpose of reporting, analysis and administration |
| Country | Name of state | Name of countries based on WHO reports |

| | | |
|---|---|---|
| Alpha-2 | Abbreviation code of the country – Two letters | Contains two letters for each state |
| Alpha-3 | Abbreviation code of the country – Three letters | Contains three letters for each state |
| latitude | Latitude of the state | |
| longitude | Longitude of the state | |
| WHO_TCC | WHO Tot. confirmed cases | Total confirmed cases WHO reports are the cumulative of confirmed cases during the time. |
| WHO_TD | WHO Tot. deaths | In WHO reports Cumulative aggregation of deaths |
| JHU_TCC | JHU Tot. confirmed cases | Cumulative of confirmed cases during the time, from JHU dataset |
| JHU_TD | JHU Tot. deaths | Cumulative aggregation of deaths in JHU dataset |
| ECDC_TCC | ECDC Tot. confirmed cases | ECDC cumulative of confirmed cases during the time |
| ECDC_TD | ECDC Tot. deaths | Cumulative aggregation of deaths in ECDC dataset |

Table 24: Joint Datasets Table of all the data sources combined.

## 4.5 Results

The comparative results of cumulative cases and deaths were visualized through Power BI software based on the collected and analyzed data from the three datasets.

A. NEGATIVE NUMBERS

After reviewing the datasets, it was discovered that the characteristics of newly confirmed cases and death cases have some negative values. These negative values were then fixed to remove inaccurate data.

The date variable in the three datasets was checked, and it was discovered that there is a one-day lag in the date variable across the various datasets. Even though the incidents occurred on January 20, WHO started reporting on 21 January 2020 for the COVID-19 cases of the previous day. The same situation was repeated on 22 January, when the statistics of 21 January were communicated with one day delay. The change occurred on 23 January, where the report included the COVID-19 cases of the same date, therefore the WHO either has no data reported for 22 January or it is aggregated with the data of January 23rd. This shows that the data has been moved on January 23 and for this reason, it should be corrected [46]. The ECDC dataset also presents the same systematic measurement error.

As mentioned earlier, for these datasets it was necessary to create a primary key combining date attribute and country code, in order to determine if public health initiatives to decrease illness severity were effective, the report's precise dates were crucial. Consequently, clinical reports with even a small error in the date variable can influence the change in the analysis explanations of the clinical data and in the wrong decision-making by the relevant authorities [46].

B. VIOLATION OF ORDER DEPENDENCY RULES

Based on logic, the total number of cases or fatalities on a given date should not be smaller than that on the preceding day. This was noticed in these datasets during the process of cumulative time series analysis, and this represents an order dependencies violation. As can be seen in Figure 28, the time series violates the Order Dependency.

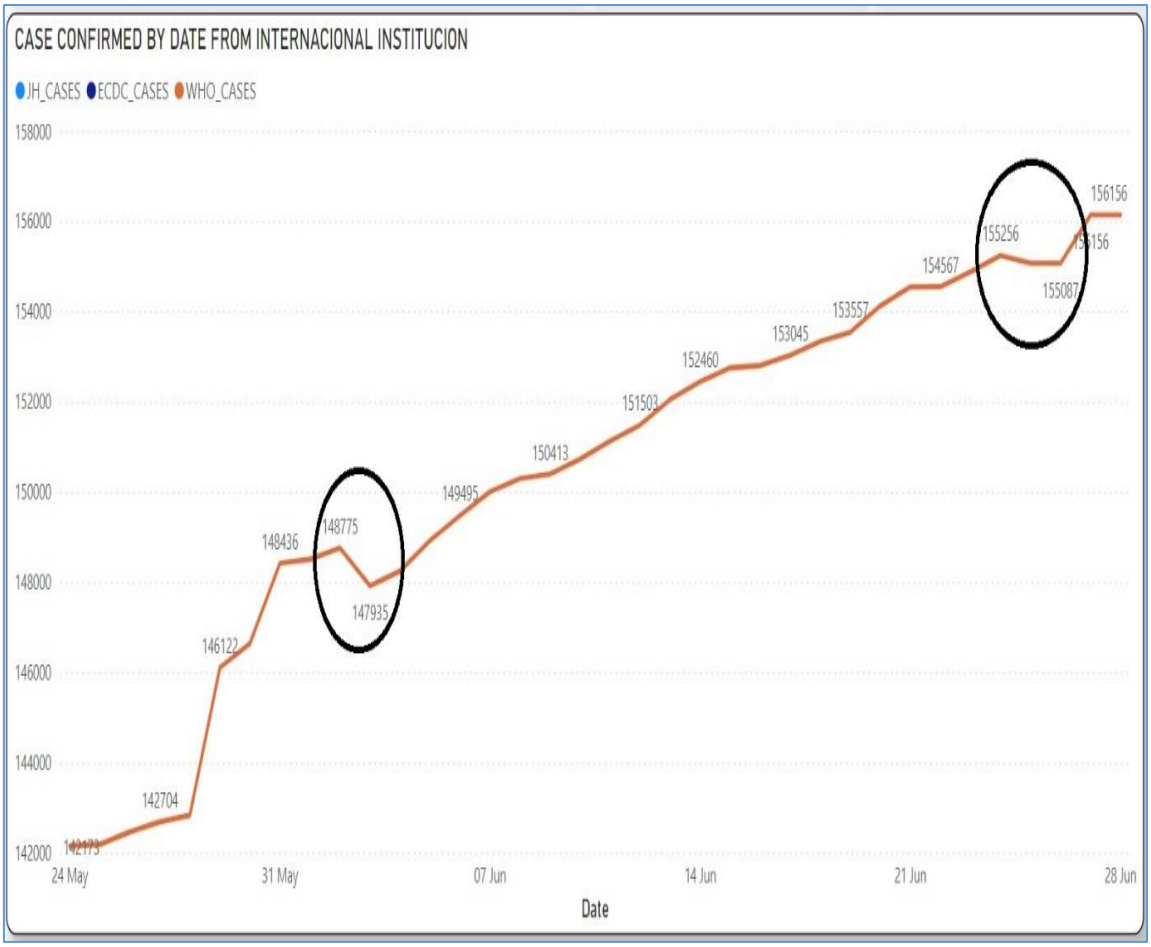



Figure 28: Order dependency violation.

C. ABNORMAL DATA POINT OR DATA PERIOD

If on one day in the cumulative time series there is an abrupt increase compared to the previous day or the subsequent day, this indicates the existence of a single abnormal point. There are a variety of reasons why this may occur, including the release of a big batch of tests or a modification to the reporting requirements, where some nations begin reporting occurrences as of a certain date.

An example where the increasing speed of cases is significantly different from other periods, called a continuous abnormal period, is presented in Figure 29.

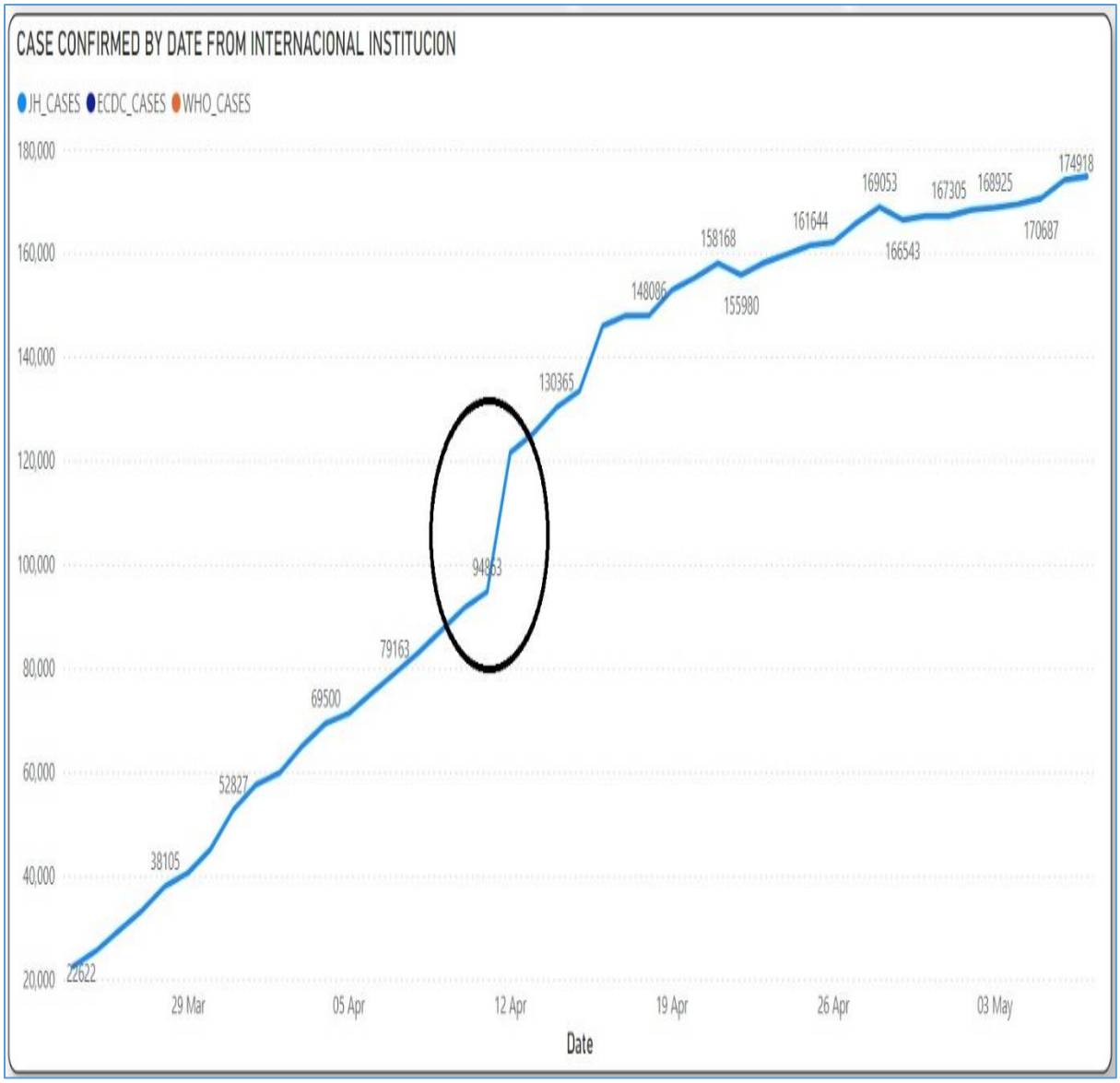



Figure 29: Single abnormal point.

### 4.5.1 Comparative Analysis of Datasets from NIPHK, WHO, JHU, ECDC

Information on the total number of COVID-19 new infection cases and total number of fatalities from the National Institute of Public Health of Kosovo (NIPHK) have been analyzed and compared with the data of the reported cases in Kosovo from the datasets of WHO, ECDC and JHU.

The difference between the cumulative number of cases reported by NIPHK and the cumulative number of Kosovo cases reported by WHO, JHU and ECDC can be seen in Figure 30.

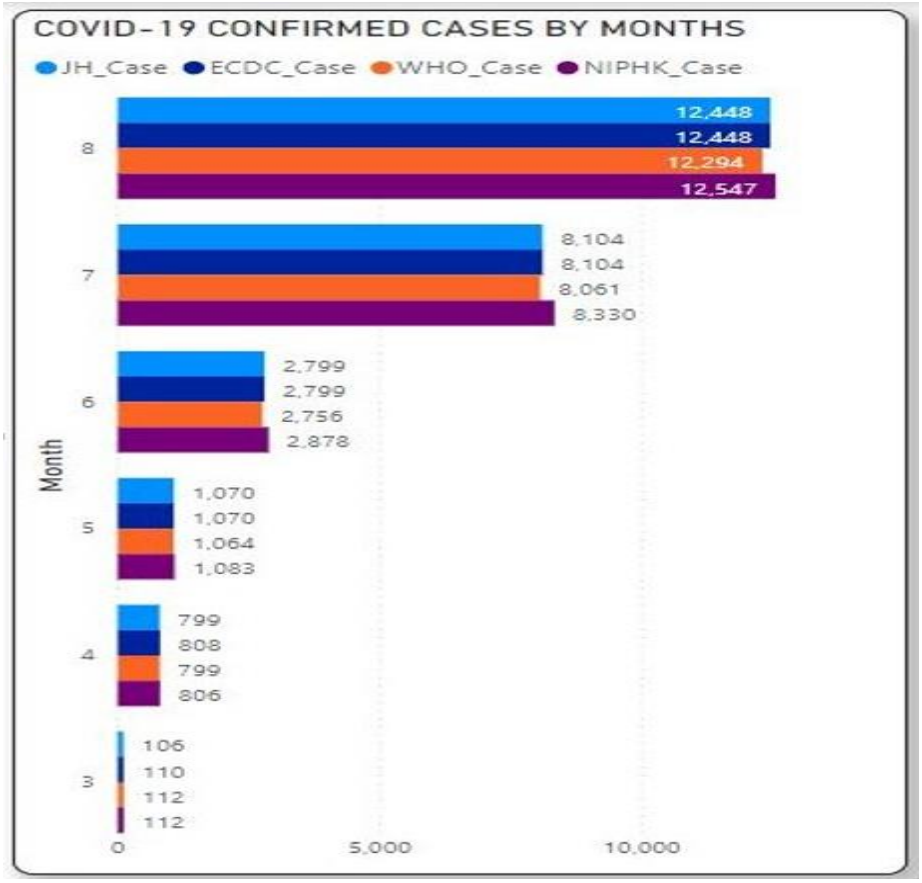



Figure 30: Cumulative confirmed case numbers from NIPHK, WHO, JHU, ECDC.

Figure 31 shows the difference in the cumulative death case numbers according to NIPHK, WHO, JHU and ECDC.

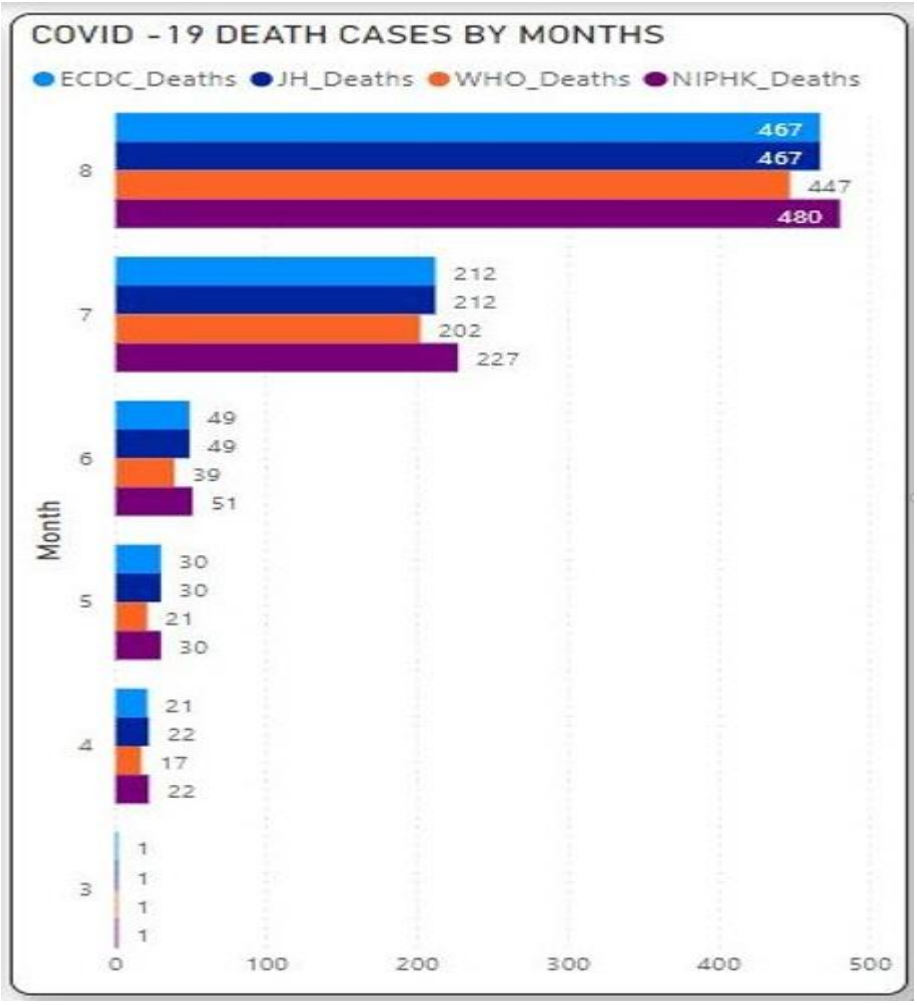



Figure 31: Cumulative death case numbers from NIPHK, WHO, JHU, ECDC.

Figure 32 displays the discrepancies for the cumulative cases of the Republic of Kosovo provided on a regular basis between the datasets of NIPHK and WHO, JHU and ECDC.

| Date | JH_CASE | ECDC_CASE | WHO_CASE | NIPHK_CASE | NIPHK vs. ECDC | NIPHK vs. JH | NIPHKvs. WHO |
|---|---|---|---|---|---|---|---|
| 31/07/2020 | 8,104 | 8,104 | 8,061 | 8,330 | ✗ | ✗ | ✗ |
| 01/08/2020 | 8,104 | 8,104 | 8,287 | 8,554 | ✗ | ✗ | ✗ |
| 02/08/2020 | 8,799 | 8,104 | 8,511 | 8,799 | ✗ | ✓ | ✗ |
| 03/08/2020 | 9,049 | 8,799 | 8,756 | 9,049 | ✗ | ✓ | ✗ |
| 04/08/2020 | 9,274 | 9,049 | 9,006 | 9,274 | ✗ | ✓ | ✗ |
| 05/08/2020 | 9,274 | 9,274 | 9,231 | 9,492 | ✗ | ✗ | ✗ |
| 06/08/2020 | 9,688 | 9,274 | 9,449 | 9,688 | ✗ | ✓ | ✗ |
| 07/08/2020 | 9,869 | 9,688 | 9,645 | 9,869 | ✗ | ✓ | ✗ |
| 08/08/2020 | 9,869 | 9,869 | 9,826 | 10,059 | ✗ | ✗ | ✗ |
| 09/08/2020 | 9,869 | 9,869 | 10,016 | 10,247 | ✗ | ✗ | ✗ |
| 10/08/2020 | 10,419 | 10,016 | 10,204 | 10,419 | ✗ | ✓ | ✗ |
| 11/08/2020 | 10,419 | 10,419 | 10,376 | 10,590 | ✗ | ✗ | ✗ |
| 12/08/2020 | 10,419 | 10,419 | 10,547 | 10,795 | ✗ | ✗ | ✗ |
| 13/08/2020 | 10,795 | 10,419 | 10,752 | 10,988 | ✗ | ✗ | ✗ |
| 14/08/2020 | 11,130 | 10,795 | 10,945 | 11,130 | ✗ | ✓ | ✗ |
| 15/08/2020 | 11,275 | 11,130 | 11,087 | 11,275 | ✗ | ✓ | ✗ |
| 16/08/2020 | 11,275 | 11,275 | 11,232 | 11,416 | ✗ | ✗ | ✗ |
| 17/08/2020 | 11,275 | 11,275 | 11,373 | 11,545 | ✗ | ✗ | ✗ |
| 18/08/2020 | 11,545 | 11,373 | 11,502 | 11,686 | ✗ | ✗ | ✗ |
| 19/08/2020 | 11,545 | 11,545 | 11,643 | 11,848 | ✗ | ✗ | ✗ |
| 20/08/2020 | 11,545 | 11,545 | 11,805 | 12,006 | ✗ | ✗ | ✗ |
| 21/08/2020 | 12,168 | 11,545 | 11,963 | 12,168 | ✗ | ✓ | ✗ |
| 22/08/2020 | 12,168 | 12,168 | 12,125 | 12,337 | ✗ | ✗ | ✗ |
| 23/08/2020 | 12,448 | 12,168 | 12,294 | 12,448 | ✗ | ✓ | ✗ |

Figure 32: Mismatches between NIPHK, WHO, JHU and ECDC.

### 4.5.2 Comparative Analysis of GRNM, WHO, JHU, ECDC Datasets

Cumulative cases of new infections and cumulative cases of deaths published by GRNM on a daily basis were analyzed and compared with WHO, ECDC and JHU datasets for reported cases for North Macedonia. The large difference in the cumulative numbers of new cases and cases of death for North Macedonia reported by GRNM, WHO, JHU and ECDC can be seen in Figure 33 and Figure 34.

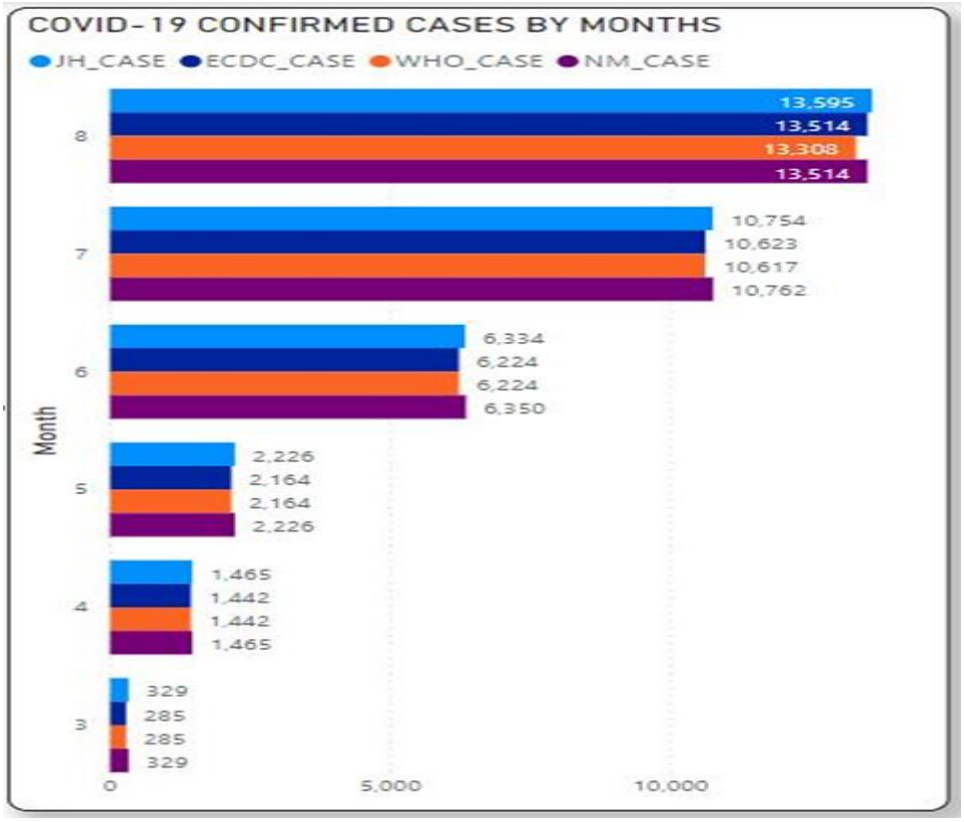



Figure 33: Cumulative confirmed case numbers from GRNM, WHO, JHU, ECDC.

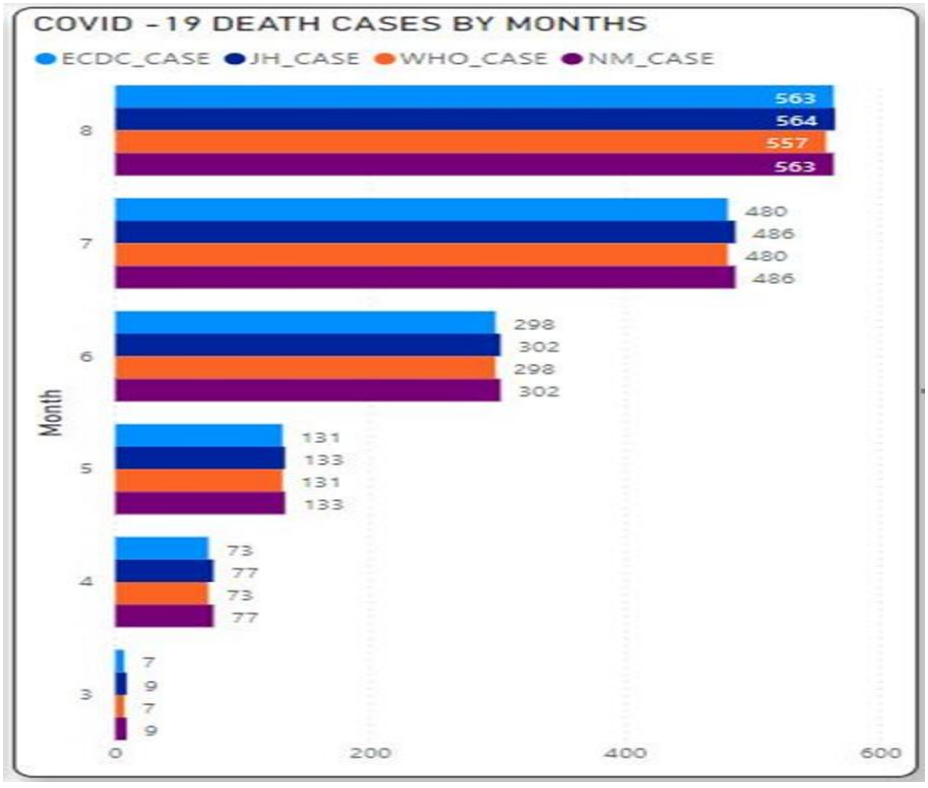
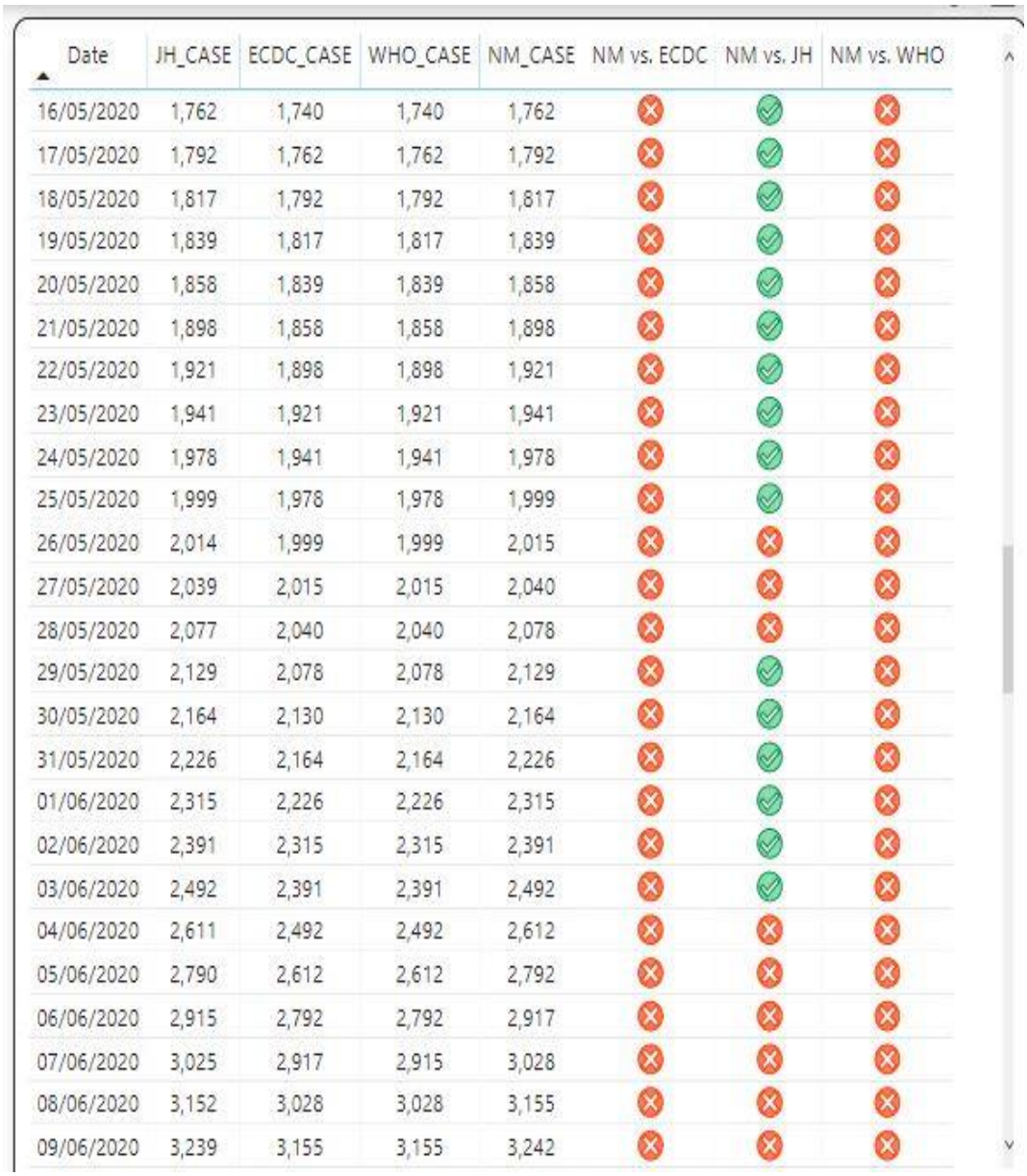


Figure 34: Cumulative death case numbers from GRNM, WHO, JHU, ECDC.

For the cumulative cases of the Republic of North Macedonia provided on a daily basis, Figure 35 displays the discrepancies between the datasets of GRNM and WHO, JHU and ECDC.



| Date | JH_CASE | ECDC_CASE | WHO_CASE | NM_CASE | NM vs. ECDC | NM vs. JH | NM vs. WHO |
|---|---|---|---|---|---|---|---|
| 16/05/2020 | 1,762 | 1,740 | 1,740 | 1,762 | ✗ | ✓ | ✗ |
| 17/05/2020 | 1,792 | 1,762 | 1,762 | 1,792 | ✗ | ✓ | ✗ |
| 18/05/2020 | 1,817 | 1,792 | 1,792 | 1,817 | ✗ | ✓ | ✗ |
| 19/05/2020 | 1,839 | 1,817 | 1,817 | 1,839 | ✗ | ✓ | ✗ |
| 20/05/2020 | 1,858 | 1,839 | 1,839 | 1,858 | ✗ | ✓ | ✗ |
| 21/05/2020 | 1,898 | 1,858 | 1,858 | 1,898 | ✗ | ✓ | ✗ |
| 22/05/2020 | 1,921 | 1,898 | 1,898 | 1,921 | ✗ | ✓ | ✗ |
| 23/05/2020 | 1,941 | 1,921 | 1,921 | 1,941 | ✗ | ✓ | ✗ |
| 24/05/2020 | 1,978 | 1,941 | 1,941 | 1,978 | ✗ | ✓ | ✗ |
| 25/05/2020 | 1,999 | 1,978 | 1,978 | 1,999 | ✗ | ✓ | ✗ |
| 26/05/2020 | 2,014 | 1,999 | 1,999 | 2,015 | ✗ | ✗ | ✗ |
| 27/05/2020 | 2,039 | 2,015 | 2,015 | 2,040 | ✗ | ✗ | ✗ |
| 28/05/2020 | 2,077 | 2,040 | 2,040 | 2,078 | ✗ | ✗ | ✗ |
| 29/05/2020 | 2,129 | 2,078 | 2,078 | 2,129 | ✗ | ✓ | ✗ |
| 30/05/2020 | 2,164 | 2,130 | 2,130 | 2,164 | ✗ | ✓ | ✗ |
| 31/05/2020 | 2,226 | 2,164 | 2,164 | 2,226 | ✗ | ✓ | ✗ |
| 01/06/2020 | 2,315 | 2,226 | 2,226 | 2,315 | ✗ | ✓ | ✗ |
| 02/06/2020 | 2,391 | 2,315 | 2,315 | 2,391 | ✗ | ✓ | ✗ |
| 03/06/2020 | 2,492 | 2,391 | 2,391 | 2,492 | ✗ | ✓ | ✗ |
| 04/06/2020 | 2,611 | 2,492 | 2,492 | 2,612 | ✗ | ✗ | ✗ |
| 05/06/2020 | 2,790 | 2,612 | 2,612 | 2,792 | ✗ | ✗ | ✗ |
| 06/06/2020 | 2,915 | 2,792 | 2,792 | 2,917 | ✗ | ✗ | ✗ |
| 07/06/2020 | 3,025 | 2,917 | 2,915 | 3,028 | ✗ | ✗ | ✗ |
| 08/06/2020 | 3,152 | 3,028 | 3,028 | 3,155 | ✗ | ✗ | ✗ |
| 09/06/2020 | 3,239 | 3,155 | 3,155 | 3,242 | ✗ | ✗ | ✗ |

Figure 35: Mismatches between GRNM, WHO, JHU and ECDC.

## 4.6 Summary

Initially, the three official COVID-19 datasets, which serve as the primary source of information for researchers worldwide, were examined in this chapter for various forms of abnormalities.

Anomalous data was discovered using the Violation of Order Dependencies approach, followed by an analysis of the weekend/holiday report delay and the identification of the abnormal data point or data period.

Therefore, it was understood that the results for scientific purposes can be significantly affected even by minor errors in the official datasets for COVID-19. Consequently, in order to achieve better scientific models and interpretations, it is crucial to consider how accurate, current, and comprehensive the COVID-19 official databases are.

Through this chapter hypothesis 2 is confirmed that the cleansing process of a complex source data in Data Quality Services (DQS) by implementing appropriate techniques and metrics will improve the data quality and may be considered as the first step before offering e-services.

In order to compare and analyze the data from these datasets, they were initially integrated into an appropriate format. Through the Power BI software, it was possible to see the difference between the WHO, JHU and ECDC datasets with the NIPHK and GRNM datasets for Kosovo and North Macedonia new cases and cumulative deaths.

# 5

# Matching and Linking Large Datasets from Multiple Resources

## 5.1 Introduction

Due to the enormous volume of data and the numerous different data sources with various data structures in government entities, it becomes quite challenging to analyze and improve the quality of data through the matching and linking process.

An essential goal in an efficient management of several data sources inside an organization is achieving good quality of the data in these data sources. If there are problems with the quality of the data or if the quality of the data is not at an appropriate level, many problems can occur in an organization. For example, incorrect decisions made by superiors due to inaccurate data, higher operating costs, and a lack of customer satisfaction are just a few of the issues that can arise [47]. When comparing and matching data from a large and different number of databases, identifying and resolving inconsistencies with the aim of providing qualitative data becomes more sensitive. Rows from several data sources that describe the same real-world object are joined in the process of matching enormous amounts of data. Data standardization, record indexing to limit the number of records to compare, row and field comparison, and identification of the rows that match or do not match, or match partly are all necessary steps in this process. The difficulty of treating volumes of the data is higher if the datasets chosen for the process of matching and linking increases exponentially in volume and quantity.

When the process of linking and matching of records is done between the two or more databases, it can be quite difficult and challenging to effectively measure the quality of the

data and evaluate the completeness dimension in data matching applications that have been studied [48].

There are two things that need to be done to assure data quality: models of data in datasets should be well specified, and values of data should be precise [49].

Data quality may be impacted by three key groups [50]:

- Data producers
- Data Custodians or protectors
- Data consumers.

Customer dissatisfaction with the data, less revenue, increased expenditures, and the need for more resources to reconcile the data are all consequences of the poor data quality.

When evaluation process related to quality of the data in those sources is completed, now the critical task is to select the proper method with the intention of producing data in high quality. Moreover, it is essential that all parties involved in the collection, processing and analysis of relevant data in datasets participate in all processes.

For the purpose of matching and linking data from various data sources, there are a variety of current algorithms that may be utilized. Levenshtein distance (LV), Damerau Levenshtein distance (DL), Longest Common Substring (LCS), Optimal String Alignment (OSA), etc. are some of the most used algorithms for this approach.

In order to use the proper data matching algorithms for linking personal records, this chapter provides first the data quality evaluation of personal records collected from numerous data sources as a prerequisite for having better quality of the data. The techniques that we will employ can be used to analyze a variety of datasets. Section 5.2 presents and discusses assessment of data quality. Data quality dimensions are presented in Section 5.3. Section 5.4 presents findings and methodology from data quality assessment. Section 5.5 treats algorithms for matching and linking personal data. The LV algorithm and the DL algorithm are suitable algorithms that can be used in circumstances where data is manually entered into electronic records and the data entry process during typing caused many spelling errors and other errors of this nature. We will analyze and compare the quality and performance of these algorithms in datasets with millions of records. We will give guidance on how to select the appropriate algorithm for matching and connecting

personal data stored in relational databases with the intention of assessing and enhancing data quality. Section 5.6 presents improving algorithms for matching and linking of personal records by adding advanced features in evaluated algorithms with the aim to reduce time and increase quality. Section 5.7 summarizes this chapter.

## 5.2 Data Quality Assessment

Data quality can be defined as a metric showing the extent to which data is appropriate for use by data consumers through the use of electronic services. The evaluation of the quality of the data is the process of deciding if the data meets the requirements necessary to complete projects or satisfy business demands, as well as whether it is of the right quantity and quality to support its intended purpose.

Identifying the quality level of the data in two datasets with more than a million rows or records each is the main goal of the data quality assessment. This step is a crucial step that can be described as a prerequisite step, before the procedure for matching and linking is carried out in these two datasets.

Data assessment is used to determine whether the data is of the right type, quantity, and quality to support its intended purpose as well as whether it satisfies the requirements for the projects or enterprises for which it is intended.

Several tasks need to be done which are classified into two phases when evaluating the quality of the data.

In order to comprehend where and how data are kept, the first phase analyzes the dataset's structure. Additionally, it indicates the group of crucial areas that are most suited for quantitative evaluation of data quality.

In the second phase, automatic software scripts are used to evaluate the data quality with a concentration on the following quality dimensions as most important and crucial:

- Completeness
- Accuracy
- Consistency.

## 5.3 Data Quality Dimensions

When assessed and used properly, dimensions of data quality directly indicate aspects that can have an impact on the general level of data quality. The procedure with the aim of defining the key data quality dimensions provides the first step in the continuation of the subsequent evaluation phase and creates the framework for various actions that improve data quality [51]. The relevance and value of the dimensions may differ between companies and institutions and types of the data as a result of their strong context dependency.
According to [52], the following are the most often utilized dimensions for demonstrating data quality:

- **Completeness**: Presents level of how much of an object should be valuable across all of its properties (For our datasets, every data field in the register should be filled up with values meaning that fields should not be null)
- **Accuracy**: Presents attributes that accurately reflect an object's real value
- **Timeliness**: It demonstrates the extent to which the data's age is appropriate for the given purpose
- **Consistency**: Demonstrates the degree to which an information item is provided in a manner that is compatible with other similar items that have information. For instance, "The Age field must contain values in the range between 0 and 130," "Person cannot marry himself," or "Person cannot die without being born first"
- **Accessibility**: Presents accessibility of the data in a given context of usage, including appropriateness of representation.

The four categories listed below can be implemented to all dimensions that have been evaluated for data quality [52]:

- **Intrinsic**: which needs to take into account all the dimensions that convey the essential quality that data should have.
- **Accessibility**: which encompasses all elements intended to convey how data is accessible to consumers
- **Contextual**: this seeks to convey the idea that data quality should be taken into account in a particular specific context

- **Representational**: which always is intended to refer to aspects of data quality that have to do with the meaning and format of the provided data.

In Table 25 below we have presented a collection of dimensions for each of four categories with specific description for every dimension that is taken under consideration.

| **Specified Category** | **DQ Dimension** | **Description** |
|---|---|---|
| Intrinsic | Accuracy Dimension | Describes when data is accurate (free of errors) and trustworthy |
| | Believability | Describes the degree of data to be considered credible and truthfulness |
| | Objectivity | Describes the objectivity of the data that have been used |
| | Reputation | Determines whether the data's contents are given careful consideration |
| Contextual | Appropriate amount | Describes how appropriate is the amount of provided information |
| | Completeness | It indicates about the range of the data displayed for treatment |
| | Relevancy | It describes how useful, relevant, or attractive the data are |
| | Value-added | Describes the indication about if the data offers a competitiveness |
| | Timeliness | Describes the age of the processed data |
| Representational | Concise representation | Describes how compactly data is represented |
| | Ease of understanding | Describes how readable, |

| | | |
|---|---|---|
| | | comprehensible, and clear the treated data is |
| | Interpretability | Describes the degree to which the significance of the data is explained |
| | Consistency | Explains the consistent use of the same format for the presentation of all data |
| Accessibility | Access security | Explains that access is secure and can be limited |
| | Accessibility | Explains how retrievable the data is to a certain extent |

Table 25: Category and Dimensions of DQ.

The following function is used to compute each of the dimensions:

*Calculated Dimension = 1- ( Nincorrect / NTotal )*

The metric for a particular desired dimension is calculated using the equation above. Value NIncorrect stands for an incorrect value, and NTotal stands for all possible values for the dimension.

For instance, the following expression can be used for computing the completeness dimension:

*Completeness dim. = 1 - (Num. of incomplete val. / tot. num. of values)*

The accuracy of the data is a further aspect that can be examined using an illustration. This dimension conveys the degree to which data are accurate and error-free [52]. The data's accuracy dimension may include one or more variables, but when the data is accurate, there is just one. In this case, the metric is determined using the below formula when we attempt to count all the units in the error:

*Accuracy dim. = 1 - (No. of val. that in error / tot. no. of values)*

Accuracy, completeness, and consistency are three important components or dimensions of data quality that will be covered in this chapter. These indicators are often utilized in assessment methods and are best suited to our goal, which is to accomplish the matching and linking of personal information across two datasets with big amount of data. Functional and accurate system cannot exist without the enforcement of data quality standards, and also data exchange between electronic systems used by the public sector is very difficult.

## 5.4 Findings and Methodology from DQ Assessment

A set of actions are included in the methodology for evaluating the data quality of a dataset. All of the actions have been divided into the following phases for the purpose of evaluating the data quality:

- In the initial or analysis phase of the data, researchers attempt to comprehend the data's schemas and conduct interviews to fully comprehend the data that are being analyzed
- Automatic Data Quality Assessment Phase, which specifies, creates, and runs various software queries over the dataset with aim to evaluate data for the consistency, completeness and accuracy.

Included are all the steps taken during the first phase of data collecting via processes and services, as well as the data collections now in use, associated management techniques, and various DQ-related issues. Examining the database architecture and schemas where the dataset stores its data was done as part of these operations.

A list of tasks completed during this phase is provided below:

- Examine and comprehend every step that was taken to create the dataset
- Evaluation of the data types being submitted into the dataset
- Define the type of metadata that is kept in the dataset
- Selection of the most pertinent crucial fields for quantitative evaluation.

Measurement of data quality using meaningful dimensions is the goal of the automated data quality assessment phase. This procedure runs entirely automatically.

Numerous database scripts (SQL queries) were created and run on the dataset utilizing the three most common criteria were used to evaluate all the data in the critical database fields that were determined to be essential during the initial phase.

With the intention of evaluating data quality through the most popular data quality dimensions, the selected essential database fields as well as the outcomes from the scripts that were performed are reported in Table 26. Scripts that were executed include, for instance:

- Scripts that count the number of rows that are missing for the essential database fields that have been identified (Name, Surname, DOB, etc.)
- Scripts that retrieve the number of records with numerical data in fields like Name, Surname etc. (in these fields we should have only letter)
- Scripts that retrieve the number of records that have symbols like.;'<> etc. in field like Name Surname (in these fields we should not have symbols)
- Scrips that retrieve the amount of records with false DOB
- Scripts that retrieve duplicate records for the specific listed fields (for instance: identifying records of citizens which have the same Name, Surname, DOB, Place of Birth, Same Name of the father and same name of the mother).

| Search describ. | Total records with data | Blank records | Records containing "///" | Records with numbers | Txt fields with symbols (/,.!@#$%^&*~":" etc. ;) | Two consonants next to each other | Date fields not in the correct format XX-XX-XXXX (-,:,.etc.) |
|---|---|---|---|---|---|---|---|
| Name | 1,724,041 | 326,007 | 221,868 | 1 | 0 | 226,237 | n/a |
| Surname | 1,816,418 | 353,473 | 135,238 | 0 | 0 | 247,526 | n/a |
| Name & surname | 1,713,046 | 134,486 | 353,473 | 0 | 0 | 36,354 | n/a |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Name or surname | 1,827,412 | 222,620 | 295,259 | 1 | 0 | 437,407 | n/a |
| DOB | 1,826,171 | not/apl. | not/apl. | not/apl. | not/apli. | not/apl. | 354872 |
| Gender | 26,765 | 2,289,805 | 0 | 0 | 0 | n/a | n/a |
| Birthplace | 1,963,081 | 353,478 | 131,991 | 265 | 0 | 0 | n/a |
| Name, surname and DOB | 1,703,632 | 298,389 | 329,854 | 0 | 0 | 24 | |
| Name, surname and birthplace | 1,713,047 | 198,895 | 126,409 | 0 | 0 | 12,975 | n/a |
| Name, surname, DOB and Birthplace | 1,703,637 | 289,478 | 6 | 0 | 0 | 10 | n/a |
| Name, surname, DOB, birthplace and name of father | 1,666,779 | 308,748 | 6 | 0 | 0 | 2 | n/a |
| Name, surname, DOB, birthplace and name of mother | 1,691,252 | 259,178 | 6 | 0 | 0 | 5 | n/a |
| Name, surname, DOB, birthplace, name of mother and father | 1,655,867 | 301,281 | 6 | 0 | 0 | 1 | n/a |

Table 26: Results after executing scripts.

Over 2.3 million entries were found in datasets based on the technique used for data quality evaluation. The dataset's average completeness rating is 0.66. The percentage of missing data for birthplace field is (19, 90%), for name field is (25, 05%), for surname field is (21,

22%), and for gender field is (98, 84%). These statistics are shown in Table 26. We should mention that there are rows in the dataset contain person who have the same name, surname, DOB, birthplace, name of father and mother, and other information. It is clear that these records are duplicates.

The automated software scripts found some entries in the dataset with accuracy-related problems. Records with numerals or other special characters in fields where only text is anticipated are just a few examples (like Name, Surname, etc.). Comparatively to records where date fields are entered incorrectly, the quantity of these rows is quite small. For instance, 15, 32% of the records in the dataset under analysis had the date of birth entered in the incorrect format (total of 354, 872 records). The format of the database columns where the dates are stored in the database is the major cause of this. Dataset stores date values as text strings rather than utilizing the proper format for storing dates.

The automated software tools found many entries in datasets with consistency problems. Examples include records in the assessed dataset with Date of Birth before year 1900 and many potential duplicate records.

## 5.5 Matching and Linking Algorithms for Personal Data

Using the appropriate algorithm to compare all records to one another and classify them into three categories or subsets, namely the set of records that are matches, the set of records that are non-matches, and the set of records that are partial-matches, is the main objective when dealing with the linking and matching process of data. This process is done once the process for evaluating the current quality of the data of the datasets that need to be treated is complete.

According to the study of the literature, several current algorithms can be utilized for matching and linking of data.

Levenshtein distance (LV) and Damerau-Levenshtein distance (DL) algorithms have been tested and used for linking and matching processes. Since they work best when data has been manually inserted into databases, these algorithms were chosen.

The two chosen algorithms for linking and matching were created using SQL language, or Structured Query Language. The results of multiple tests were examined in order to choose the best algorithm.

### 5.5.1 Used Variables in Algorithms

The following variables were defined before using the chosen algorithm:

- “Variables considered for Matching and Linking” - We will utilize other personal identities that are present in the database as variables, such as name, surname, DOB, and gender, because our dataset lacks having a single personal identity that may be used as a linking variable, such as a personal identification number
- “Blocking variables” - Variables that do not compare rows that may have the lowest likelihood of matching, a process which reduces the amount of space that has to be searched between two datasets. By establishing comparison pairs that have a high likelihood of being matched—for example, those who were born on specified dates or other criteria—the search space is meant to be significantly reduced.

The Levenshtein distance is initially applied to two datasets that we have decided to apply. New created table in the database is used to keep logs generated from the outcome of the algorithm and contains info presented in percentage for the activities done. A signal that data is missing (the database field for that variable is empty or null) is when the value of a matching and linking variable is zero.

When comparing the two datasets for the first time, the algorithm generates three distinct sorts of results based on the chance that the records in question pertain to the same entity at different levels:

- The term "matches" is used to describe situations when there is a 100% match, or when all of the data fields from the several records under consideration completely correlate
- When data fields from several records do not line up exactly but do so with 80% to 99% accuracy, these instances are referred to as "potential matches"

- When data fields from various records do not entirely match and the discrepancies are substantial, the matching is deemed to be between 0% and 79%; these situations are referred to as "no-matches".

Attributes or fields that are used while applying the algorithm for matching and linking of personal data are listed below:

- Name
- Surname
- DOB
- Birthplace
- Father’s Name
- Father’s Surname
- Mother’s Name
- Mother’s Surname.

In order to design and implement the matching and linking algorithm, we have used a combination of blocking variables. First blocking variable thought to be the DOB. Space that it is needed to be searched was quickly decreased through comparing mainly between those that have the potential to be matched using pairs, which was accomplished by using appropriate blocking variable to split the collection of records for matching and linking into various groups (people born on the same day etc.).

After datasets have been compared, the matched records should be linked together and kept aside in specific table with DOB acting as blocking variable.

Name fiend should be used as the next blocking variable, followed by Surname field, if the remaining datasets are still too large.

### 5.5.2 Results form using Matching and Linking Algorithms

To execute algorithms it was needed to create a high performance hardware infrastructure. Testing infrastructure properties and volumes of data that are compared are shown in table below:

| No1. | Testing infrastructure - hardware | |
|---|---|---|
| 1. | RAM Memory | 128 GB |
| 2. | HDD | 14 TB |
| 3. | Processor | Intel® Xeon® Platinum 8164 2.00 GHz (16 CPUs) |
| | Datasets Volume | |
| 1. | Dataset 1 | 2.5 Million Rows |
| 2. | Dataset 2 | 1.85 Million Rows |

Table 27: Hardware infrastructure – testing environment.

The performance comparison of the two algorithms is shown in Table 28. The Levenshtein distance (LV) algorithm was significantly quicker than the Damerau-Levenshtein distance (DL) algorithm, as seen in the table. Performance or speed indicator is vital and very important when we do the process for comparing millions of records in our datasets with each other.

In order to go further with the matching and linking exercise, because of speed reason it was decided to use the Levenshtein distance (LV) algorithm. We will compare the data in the designated fields and provide a similarity percentage.

| | **Damerau Levenshtein Distance** | **Levenshtein Distance** |
|---|---|---|
| How long does it take to execute the algorithm for 1 record of dataset1 on 1 record in dataset2? | 0 seconds | 0 seconds |
| How long does it take to execute the algorithm for 1 record of dataset1 on 10 | 0 seconds | 0 seconds |

| | | |
|---|---|---|
| records in dataset2? | | |
| How long does it take to execute the algorithm for 1 record of dataset1 on 100 records in dataset2? | 3 seconds | 0 seconds |
| How long does it take to execute the algorithm for 1 record of dataset1 on 1000 records in dataset2? | 39 second | 1 seconds |

Table 28: Comparison of the two algorithms' performance for linking and matching.

Using the date of birth as a blocking variable, through Table 29 we compared the results in percentage form.

| | **Number of matching** | **Total %** |
|---|---|---|
| 100% | 47,606 | 0.01% |
| 90%-99.99% | 805,765 | 0.20% |
| 80%-89.99% | 992,109 | 0.25% |
| 70%-79.99% | 595,112 | 0.15% |
| 60%-69.99% | 503,831 | 0.13% |
| 50%-59.99% | 1,383,970 | 0.35% |
| 40%-49.99% | 14,731,744 | 3.69% |
| 30%-39.99% | 108,451,322 | 27.19% |
| 20%-29.99% | 218,032,275 | 54.65% |
| 10%-19.99% | 53,387,296 | 13.38% |
| 0%-9.99% | - | 0.00% |
| Total | 398,931,030 | 100.00% |

Table 29: Compared Results.

The outcome of the process is displayed in Table 30 after the following stage of removing all duplicate data for the percentage matching higher than 50%.

| **Range of matching in %** | **Number of matching** | **Total %** |
|---|---|---|
| 100% | 32,929 | 1.26% |
| 90%-99.99% | 528,043 | 20.20% |
| 80%-89.99% | 539,429 | 20.63% |
| 70%-79.99% | 270,601 | 10.35% |
| 60%-69.99% | 296,158 | 11.32% |
| 50%-59.99% | 947,449 | 36.24% |
| Total | 2,614,609 | 100% |

Table 30: After removing duplicate data, results for a range greater than 50%.

The source code used to generate results using the Levenshtein distance algorithm is shown in APPENDIX A.

## 5.6 Improving Algorithms for Matching and Linking of Personal Records by Adding Weight Feature

Based on the needs from researchers many new features in algorithms and functions are added with the aim that the result will be more accurate.

When we implemented the Levenshtein algorithm in our personal records datasets, we noticed that all the fields have the same weight or importance. One example is shown in table below:

| FIRST NAME | LAST NAME | DOB | BITHPLACE | FATHER 'S FIRST NAME | FATHER'S LAST NAME | MATHER' S FIRST NAME | MATHER'S LAST NAME | MATCHING % |
|---|---|---|---|---|---|---|---|---|
| AZEM | MAXHUNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 14.28% | 14.28% | | 14.28% | 14.28% | 14.28% | 14.28% | 14.28% | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A**D**EM | MA**GJ**UNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 10.71 | 10.20% | | 14.28% | 14.28% | 14.28% | 14.28% | 14.28% | **92.31%** |

Table 31: Personal records with same weight for all columns.

As we can see in this example, the percentage of mistakes is the same no matter if the mistake is in the fiend First Name or in the field Mathers's First Name because all the fields have 14.28 % in total percentage.

For the datasets that we compared, some fields are more important that the other to show the accuracy of the records.

For the same record compared as in Table 32, we can notice that when we add feature Weight or Importance for every column, we have different results in percentage.

One example is shown in table below:

| FIRST NAME | LAST NAME | DOB | BITHPLACE | FATHER 'S FIRST NAME | FATHER'S LAST NAME | MATHER' S FIRST NAME | MATHER'S LAST NAME | MATCHING % |
|---|---|---|---|---|---|---|---|---|
| AZEM | MAXHUNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 20% | 20% | | 14% | 11% | 12% | 11% | 12% | |
| | | | | | | | | |
| A**D**EM | MA**GJ**UNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 15% | 14% | | 14% | 11% | 12% | 11% | 12% | **89.28%** |

Table 32: Personal records with different weight for specific columns.

In the table below, we can see the difference in percentage when we compare the results of matching two records with the same mistake but with different weight or importance of columns compared.

| **FIRST NAME** | **LAST NAME** | **DOB** | **BITHPLACE** | **FATHER 'S FIRST NAME** | **FATHER'S LAST NAME** | **MATHER 'S FIRST NAME** | **MATHER' S LAST NAME** | **MATCHING %** |
|---|---|---|---|---|---|---|---|---|
| AZEM | MAXHUNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 14.28% | 14.28% | | 14.28% | 14.28% | 14.28% | 14.28% | 14.28% | |
| | | | | | | | | |
| A**D**EM | MA**GJ**UNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 10.71 | 10.20% | | 14.28% | 14.28% | 14.28% | 14.28% | 14.28% | **92.31%** |
| **FIRST NAME** | **LAST NAME** | **DOB** | **BITHPLACE** | **FATHER 'S FIRST NAME** | **FATHER'S LAST NAME** | **MATHER 'S FIRST NAME** | **MATHER' S LAST NAME** | **MATCHING %** |
| AZEM | MAXHUNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 20% | 20% | | 14% | 11% | 12% | 11% | 12% | |
| | | | | | | | | |
| A**D**EM | MA**GJ**UNI | 15.12.1985 | GJILAN | HASAN | MAXHUNI | HAVE | MAXHUNI | |
| 15% | 14% | | 14% | 11% | 12% | 11% | 12% | **89.28%** |

Table 33: Comparing personal records with same and different weight for columns.

The comparison outcome is presented in percentage form in Table 34, with DOB field as blocking variable.

| | **Number of matching** | **Total %** |
|---|---|---|
| 100% | 47,606 | 0.01% |
| 90%-99.99% | 965,234 | 0.24% |
| 80%-89.99% | 1,142,298 | 0.27% |
| 70%-79.99% | 834,892 | 0.20% |
| 60%-69.99% | 789,269 | 0.19% |
| 50%-59.99% | 1,751,257 | 0.43% |
| 40%-49.99% | 17,648,285 | 4.42% |
| 30%-39.99% | 115,287,487 | 29.00% |
| 20%-29.99% | 214,010,278 | 53.55% |
| 10%-19.99% | 46,457,424 | 11.71% |
| 0%-9.99% | - | 0.00% |
| Total | 398,931,030 | 100.00% |

Table 34: Compared Results with different weights of fields.

The outcome is presented in Table 35 after the following stage of removing all duplicate values for the percentage matching higher than 50%.

| **Range of matching in %** | **Number of matching** | **Total %** |
|---|---|---|

| | | |
|---|---|---|
| 100% | 32,929 | 1.26% |
| 90%-99.99% | 592,124 | 22.71% |
| 80%-89.99% | 580,289 | 22.24% |
| 70%-79.99% | 291,357 | 11.15% |
| 60%-69.99% | 346,159 | 13.19% |
| 50%-59.99% | 771,751 | 29.45% |
| Total | 2,614,609 | 100% |

Table 35: After removing duplicate data, results for a range greater than 50%.

The improved source code used to generate results using Levenshtein Distance Algorithm is shown in APPENDIX B.

## 5.7 Improving Algorithms for Matching and Linking of Personal Records by comparing similar letters in Albanian alphabet

During the process of analyzing the data that has been compared for the purpose of matching the data from different datasets, we noticed that in the Albanian language there are some letters that are similar or that are often used when writing names, surnames, birthplaces and other important fields when filling in the citizens' data. For example, it is often wrong when writing the name Qerim when this name is written as Çerim. Both of these letters in the Albanian language have the same pronunciation but are used in specific cases. Through the improvement of Levenstein's algorithm, we managed to reduce the distance from 1 to 0.6, for such errors and other errors listed in the table below, while the distance for other letters that are not in the table 36 the distance is 1.

| **Number.** | **First comparative letter** | **Second comparative letter** |
|---|---|---|
| 1. | “e” | “ë” |
| 2. | “ë” | “e” |

| | | |
|---|---|---|
| 3. | “i” | “j”, “y” |
| 4. | “j” | “i”,”y” |
| 5. | “y” | “i”,”j” |
| 6. | “q” | “ç” |
| 7. | “ç” | “q” |

Table 36: Similar letters in Albanian Alphabet.

Part of the code is shown as follows. The all source code used to generate results using improving the Levenshtein Distance Algorithm for similar letters in the Albanian alphabet is shown in APPENDIX C.

Similar letters in the Albanian alphabet.

```
1:  USE [DBIMPROVED]
2:  GO
3:  /****** Object:  UserDefinedFunction
    [dbo].[LevenshteinDistance]******/

4:  SET ANSI_NULLS ON
5:  GO
6:  SET QUOTED_IDENTIFIER ON
7:  GO
8:  --  SELECT [dbo].[neighbors] ('a', 'q' )
9:  Create FUNCTION [dbo].[neighbors](@s1 nvarchar(100), @s2
    nvarchar(100))
10: RETURNS bit
11: AS
12: BEGIN
13: declare @c bit
14: set @c=case when @s1 in ('e') and @s2 in ('ë') then 1 else
15:        case when @s1 in ('ë') and @s2 in ('e') then 1 else
16: case when @s1 in ('i') and @s2 in ('j','y') then 1 else
17: case when @s1 in ('j') and @s2 in ('i','y') then 1 else
18: case when @s1 in ('y') and @s2 in ('i','j') then 1 else
19: case when @s1 in ('q') and @s2 in ('ç') then 1 else
20. case when @s1 in ('ç') and @s2 in ('q') then 1 else
```

```
21.     0
22.     end end end end end end end
23.     RETURN @c
24.     END
```

Table 37 displays the outcome of putting the stated changes into practice for the matching higher than 50%.

| Range of matching in % | Number of matching | Total percentage |
|---|---|---|
| 100% | 32,929 | 1.26% |
| 90%-99.99% | 622,458 | 23.92% |
| 80%-89.99% | 599,954 | 23.03% |
| 70%-79.99% | 334,478 | 12.84% |
| 60%-69.99% | 379,127 | 14.57% |
| 50%-59.99% | 645,663 | 24.80% |
| Total | 2,614,609 | 100% |

Table 37: Results for range more than 50% after implementing improvements for similar letters in Albanian alphabet.

By implementing an improved Levenshtein algorithm when similar letters have less distance in the Albanian alphabet, we reduce the time when comparing data from different data sets and we come to better results in the faster way.

## 5.8 Improving Algorithms for Matching and Linking of Personal Records by specifying distance of edit operations

By applying Levenshtein's approach, the distance between two documents is estimated as the least amount of changes that need to be done in order to create one document from another document.

The editing techniques used by this algorithm are listed below:

- Insert
- Delete
- Substitute.

By implementing all three described operations into practice, it is possible to create a document from another document by changing, removing, or adding a certain number of characters.

During the implementation of this algorithm, we improved the definition of the importance (distance) of editing operations, where we assigned the distance 0.6 to the substitution operation, 1.2 to the Insert operation, and 1.2 to the Delete operation.

Part of the source code is shown as follows. The complete source code that is used to generate results using improved Levenshtein Distance Algorithm for edit operations is shown in APPENDIX D.

Improving Levenshtein Distance Algorithm for edit operations.

```
1:  BEGIN
2:  if (len(@s2) = 0)
3.      begin
4:      return @s2
5:  end
6:  DECLARE @s1_len int, @s2_len int, @i int, @j int, @s1_char
    nchar, @c int, @c_temp float, @n bit, @nk int, @k int,
        @cv0 varbinary(8000), @cv1 varbinary(8000)
7:  SELECT @s1_len = LEN(@s1), @s2_len = LEN(@s2), @cv1 =
    0x0000, @j = 1, @i = 1, @c = 0, @nk = 0, @k = 1
8:  WHILE @j <= @s2_len
9:  SELECT @cv1 = @cv1 + CAST(@j AS binary(2)), @j = @j + 1
10: WHILE @i <= @s1_len
11: BEGIN
12: SELECT @s1_char = SUBSTRING(@s1, @i, 1), @c = @i, @cv0 =
    CAST(@i AS binary(2)), @j = 1
13: WHILE @j <= @s2_len
14: BEGIN
```

```
15:     SET @c = @c + 1
16:     SET @c_temp = (CAST(SUBSTRING(@cv1, @j+@j-1, 2) AS int) +
                CASE WHEN @s1_char = SUBSTRING(@s2, @j, 1) THEN 0
        ELSE 1 END)
17:     IF @c > @c_temp SET @c = @c_temp
18:     SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j+1, 2) AS int)+1
19:     IF @c > @c_temp SET @c = @c_temp
20.     SELECT @cv0 = @cv0 + CAST(@c AS binary(2)), @j = @j + 1
21.     END
22.     SELECT @cv1 = @cv0, @i = @i + 1
23.     declare @ins int =0, @del int = 0
24.     if @s1_len > @s2_len
25.     begin
26.     set @del = @s1_len - @s2_len
27.     end
28.     else
29.     begin
30.     set @ins =  @s2_len - @s1_len
31.     end
32.     END
33.     RETURN  (@c * 0.6) +  (@ins * 1.2) +  (@del *1.2) - (@ins
        * 0.6) - (@del * 0.6)
34      END
```

By implementing an improved Levenshtein algorithm when distance in edit operations is specified, we reduce the time when comparing data from different data sets and we come to better results in a faster way.

## 5.9 Summary

In this study, through data quality dimensions, we evaluated the datasets of personal records. Datasets containing personal records are assessed based on criteria including consistency, accuracy, and completeness.

After evaluating the datasets that contain personal data for data quality, we implemented various data matching algorithms to link personal records in these datasets. We concluded

that the Levenshtein Algorithm is the best algorithm for data matching and linking personal records based on quality and performance.

We also added new features to algorithms that treat data from multiple resources with the aim to improve quality of data in order to reduce the time needed to obtain the required result.

Through this chapter hypothesis 4 is confirmed that applying proper algorithms to evaluate the approaches chosen for matching and linking huge datasets of personal records from various sources can increase the quality of the data.

Also, through this chapter hypothesis 1 is confirmed that the harmonization process of the data will improve data quality outputs compared from multiple resources containing the same nature of the data.

The findings of this study demonstrate the significance of evaluating the data quality as the first step before delivering electronic services and also it demonstrates the significance of selecting the appropriate algorithm for linking and matching of the data from multiple resources.

# 6

# E-services Evaluation and Delivery Model Using Data Cleansing Logical Constraints

## 6.1 Introduction

Effective electronic services in government portals can be achieved by increasing the quality of data to a high level, considering that these data are retrieved from multiple data sources. Also, the government as a strategy and main goal should have the delivery of services in one stop shop by increasing the quality of the data.

Providing public services is made possible through government e-services as an online channel in real time 24 hours a day, 7 days a week. Therefore, providing quality services is possible through adequate and qualitative data in order not to have a negative impact on the trust of citizens and the lack of citizen's satisfaction towards these electronic services.

E-services can be defined as electronic applications that enable citizens to use various online services without depending on the presence of government officials [70].

Providing e-services in itself involves the merging of data of the same entity from different data sources, so-called as an integrated process of e-Services.

The implementation of e-Government through e-services ensures efficiency of public services as well as transparency and accountability of the Government to the citizens [71].

High quality data can be characterized as data that are fit for use in electronic services by citizens [48].

In research on public administration, the use of electronic services represents an important topic for the entire public sector [72].

In order to improve transparency and accountability, citizens and businesses are provided with efficient and effective access to government services through information technology [73].

Depending on which services are digitized, but in general e-Government services include access to forms and services of different sectors, access to information on development policies, various government plans, employment opportunities, transparent procurement and election information [74].

By reducing government bureaucratic transactions and the saving of various costs, some models of electronic government were adopted, which resulted in the increase of the ability to fulfill the satisfaction of the citizens [75].

In addition, in the Kosovo Government portal, the quality of services as an important objective depends directly on the prerequisite of the evaluation of e-Services.

With the purpose of registering various data in electronic registers, after the war of 1999, the Republic of Kosovo created new electronic registers and for which the care for the data quality was not sufficient, allowing the registration of data without any criteria or dimensions.

These datasets came from different sources, with different structures, with a large volume of data and with a very low data quality, where as a result the assessment and improvement of the data quality represents a very difficult task with the aim of delivering quality e-Services.

Through this chapter, for the datasets of personal and vehicle registration data, the data quality dimensions are first determined and then is performed assessment of electronic services provided by Kosovo Government.

The analysis of the datasets mentioned above is performed using data quality dimensions such as Accuracy, Completeness, Consistency, Uniqueness and Timeliness.

With the aim of providing the best possible G2C (Government to Citizens) services through the government portal, the impact of the implementation of data quality standards and dimensions will also be assessed.

In addition, the greatest importance will be given to the assessment and improvement of:

- DQ by specifying dimensions

- Quality of system by improving electronic systems
- E-Services quality through Government portal.

Using data quality dimensions, microservices and the integration of services in the Kosovo Portal (e-Kosovo) for data with multiple data sources, it is intended to provide the model of Government e-Services with improved data quality.

The chapter is organized as follows starting with introduction in section 6.1. In section 6.2, we will present methodology and challenges in data quality improvement in imposing E-services for Kosovo Government Institutions. In section 6.3, we will present logical constraints to ensure data quality in the Personal Documents Register (PDR) and Vehicle Register (VR). In section 6.4, we will present outcomes from assessment and improvement of data quality on source electronic registers. In section 6.5, we will present a model for integrating e-Services in electronic platforms – e-Kosovo Model. Section 6.6 summarizes this chapter.

## 6.2 Methodology and Challenges in DQ Improvement in Imposing E-Services for Kosovo Government Institutions

Data quality represents the level of data suitable for using by potential data consumers. In order to decide that certain data are intended to be used for different business processes, these data must first be assessed in terms of their quality, as a prerequisite for providing the best possible e-services.

Some of the obstacles listed below can limit the implementation speed of government e-services [76]:

- Transformation speed- transformation through different phases takes time due to various resistances and problems, including changes in employee procedures, insufficient budget, legal issues, awareness, training, etc.
- Implementing transactions for Government to Citizens (G2C) process - this category of e-services includes larger and more complicated transactions, unlike other categories such as G2B and G2G
- Issues related to Security and Privacy - Government e-Services must be subject of

personal information that can be misused, being protected using various adequate procedures, methods and techniques.

## 6.3 Logical Constraints to Ensure DQ in the Personal Documents Register (PDR) and Vehicle Register (VR)

Data quality characteristics such as dimensions or logical constraints represent the overall level of data quality based on their measurement and use in an adequate manner. One of the important steps of data assessment and improvement is the process of identifying the relevant dimensions with the aim of providing good e-services in Government Institutions [52].

Immediately after the 1999 war, due to the needs of various institutions, different source registers were created by the Kosovo government on platforms such as Microsoft Excel, Access, but the quality of the data was not at an adequate level.

The management, which includes the assessment and improvement of the data quality, in the Personal Document Register (PDR) and the Vehicle Register (VR), largely determines the quality of government e-services for businesses and citizens, since these two electronic registers contain the largest volume of general e-services transactions.

Ensuring good e-services is achieved by preventing the input of incorrect data and other anomalies in the source registers, which in our case are the PDR and VR registers. This objective is achieved through the process of data analysis and their improvement through adequate activities and actions.

There is a list below of dimensions that will be used to assess and improve the data quality of PDR and VR registers:

- Accuracy
- Completeness
- Consistency
- Uniqueness
- Timeliness.

The accuracy represents the level of correctness and reliability of the data [64]. Concretely, the accuracy shows to what level of reality the information stored in the database corresponds, determining if the data values are correct for a specific object.

The following list presents some of the data quality indicators that should be checked at least once a month in the PDR and VR registers with the aim of continuing the data quality accuracy process:

- Citizens with date of marriage later than the date of death
- Citizens whose age difference is less than 13 years with the age of their parents
- Citizens with multiple spouses
- Citizens with more than 20 registered children
- Citizens containing more than two parents in parents field
- Registrations containing date of registration later than date of expiration
- Registrations with validity period more than one year
- Physical owners having more than 10 active vehicles registered
- Vehicles having more than one active owner
- Vehicles without an active owner.

If such findings are identified, adequate actions must be taken to correct them in the PDR and VR databases.

The completeness represents the level of description of the set of real-world objects through a certain collection of data [57]. Concretely, the completeness dimension shows the level and amount of characteristics that exist in the database data, where the missing data should be collected.

The following list presents some of the data quality indicators that should be checked at least once a month in the PDR and VR registers with the aim of continuing the data quality completeness process:

- Citizens that contain null value in name field
- Citizens that contain null value in surname field
- Citizens that contain null value in PIN field

- Citizens that contain null value in DOB field
- Citizens that contain null value in gender field
- Citizens that contain null value in citizenship field
- Citizens that contain null value in municipality of dwelling fields
- Vehicles that contain null value in license plate field
- Vehicles that contain null value in ChassisNo field
- Physical and legal owners that don't contain unique ID
- Vehicles that contain null value in registration and expiration date field
- Vehicles with uncompleted data for model and country of origin
- Vehicles registered without defining registration municipality.

If such findings are identified, adequate actions must be taken to correct them in the PDR and VR databases.

Data consistency represents the correctness and completeness of the logical relationship between correlated data [1]. Concretely, the consistency of the data shows if all the registers and systems understand and interpret the data in a similar way, therefore there is no deviation in the meaning of the same data.

The following list presents some of the data quality indicators that should be checked at least once a month in the PDR and VR registers with the aim of continuing the data quality consistency process:

- Fields like name and surname containing invalid characters
- Invalid format of date field for DOB, death and marriage
- Date field used for DOB, death and marriage containing non-existing datetime
- PIN is not in compliance with the rules
- Data in place of birth field or place of residence field does not comply with rules or refers to the non-existing address
- Codes in gender field, citizenship field or marital status field does not comply with rules
- Records containing not allowed characters in the TagNo and ChassisNo fields

- Records containing invalid format of TagNumber
- Vehicle registrations with invalid date of registration and expiration.

If such findings are identified, adequate actions must be taken to correct them in the PDR and VR databases.

The unique dimension of data represents the attribute of singularity of data or records, making it impossible to store the same facts more than once, which means that in the database there should not exist duplicate registration of specific data within a record.

The following list presents some of the data quality indicators that should be checked at least once a month in the PDR and VR registers with the aim of continuing the data quality uniqueness process:

- Multiple citizens having same PIN
- Multiple places and dates of the birth of the same citizen
- Multiple resident places of the citizen
- Multiple vehicles with same chassis and plates
- Owners who registered at same period of time the same vehicle
- Physical and legal owners that are not unique in the system and have active registered vehicles.

If such findings are identified, adequate actions must be taken to correct them in the PDR and VR databases.

The timeliness represents the level of data on the representation of the current situation and if they exist at the time when they are needed, with the purpose of avoiding negative impact on reports and decision-making because of late data entries and updates.

The following list presents some of the data quality indicators that should be checked at least once a month in the PDR and VR registers with the aim of continuing the data quality timeliness process:

- Citizens that have more than 110 years
- Citizens that contain invalid documents and death registration is not done
- Citizens without official place of residence

- Vehicles not registered in last 10 years
- Vehicles with owner older than 90 years
- Vehicles older than 30 years and have not been registered as old-timer

If such findings are identified, adequate actions must be taken to correct them in the PDR and VR databases.

## 6.4 Outcomes of DQ Assessment and Improvement on Core Electronic Registers

In order to assess and improve the quality of the data in the treated datasets, a particular assessment technique is used with the aim to provide most representative results. This approach entails a series of tasks that need to be done. What is utilized as a precondition is automatic data quality evaluation and improvement. To treat data to be in compliance with the five most crucial aspects that are needed to be fulfilled in order for the specific data to be considered as adequate in aspect of DQ, it is necessary that several queries on the datasets be conducted.

Measuring the quality of data collections along the most crucial DQ dimensions is the goal of automated data quality assessment and improvement.

A variety of programming scripts (SQL queries) were built and executed on the datasets, both PDR and VR, with the goal of evaluating and improving DQ of the data in specified datasets.

The following tables lists the essential fields in databases that were found and treated as crucial for DQ assessment and improvement. The results of all the scripts that were run to assess the data's quality using the most crucial DQ criteria, such as accuracy, completeness, consistency, uniqueness, and timeliness, are also displayed here.

The following table displays the results from implementing the accuracy dimension in PDR and VR datasets:

| Implementation of accuracy dimension | Value |
|---|---|

| | |
|---|---|
| Citizens with date of marriage later than the date of death | 21 |
| Citizens whose age difference is less than 13 years with the age of their parents | 226 |
| Citizens with multiple spouses | 0 |
| Citizens with more than 20 registered children | 1 |
| Citizens containing more than two parents in parents field | 0 |
| Registrations containing date of registration later than date of expiration | 100 |
| Registrations with validity period more than one year | 88 |
| Physical owners having more than 10 active vehicles registered | 92 |
| Vehicles having more than one active owner | 58 |
| Vehicles without an active owner | 32 |

Table 38: Results of implementing accuracy dimension.

The following table displays the results from implementing the completeness dimension in PDR and VR datasets:

| **Implementation of completeness dimension** | **Value** |
|---|---|
| Citizens that contain null value in name field | 74 |
| Citizens that contain null value in surname field | 100 |
| Citizens that contain null value in PIN field | 0 |
| Citizens that contain null value in DOB field | 0 |
| Citizens that contain null value in gender field | 10 |
| Citizens that contain null value in citizenship field | 2 |

| | |
|---|---|
| Citizens that contain null value in municipality of dwelling fields | 482 |
| Vehicles that contain null value in license plate field | 100 |
| Vehicles that contain null value in ChassisNo field | 58 |
| Physical and legal owners that don't contain unique ID | 71 |
| Vehicles that contain null value in registration and expiration date field | 125 |
| Vehicles with uncompleted data for model and country of origin | 78 |
| Vehicles registered without defining registration municipality | 42 |

Table 39: Results of implementing completeness dimension.

The following table displays the results from implementing the consistency dimension in PDR and VR datasets:

| **Implementation of consistency dimension** | **Value** |
|---|---|
| Fields like name and surnames containing invalid characters | 42 |
| Invalid format of date field for DOB, death and marriage | 0 |
| Date field used for DOB, death and marriage containing non-existing datetime | 37 |
| PIN is not in compliance with the rules | 25 |
| Data in place of birth field or place of residence field does not comply with rules or refers to the non-existing address | 62 |
| Codes in gender field, citizenship field or marital status field does not comply with rules | 0 |
| Records containing not allowed characters in the TagNo and ChassisNo fields | 125 |

| | |
|---|---|
| Records containing invalid format of TagNumber | 10 |
| Vehicle registrations with invalid date of registration and expiration | 38 |

Table 40: Results of implementing consistency dimension.

The following table displays the results from implementing the uniqueness dimension in PDR and VR datasets:

| **Implementation of uniqueness dimension** | **Value** |
|---|---|
| Multiple citizens having same PIN | 21 |
| Multiple places and dates of the birth of the same citizen | 40 |
| Multiple resident places of the citizen | 12 |
| Multiple vehicles with same chassis and plates | 17 |
| Owners who registered at same period of time the same vehicle | 25 |
| Physical and legal owners that are not unique in system and have active registered vehicles | 0 |

Table 41: Results of implementing uniqueness dimension.

The following table displays the results from implementing the timeliness dimension in PDR and VR datasets:

| **Implementation of Timeliness dimension** | **Value** |
|---|---|
| Citizens that have more than 110 years | 559 |
| Citizens that contain invalid documents and death registration is not done | 410 |
| Citizens without official place of residence | 28 |
| Vehicles not registered in last 10 years | 25 |
| Vehicles with owner older than 90 years | 49 |

| | |
|---|---|
| Vehicles older than 30 years and have not been registered as old-timer | 32 |

Table 42: Results of implementing timeliness dimension.

Datasets containing millions of records each were used for the assessment and improvement process of data quality.

The results table shows that a lot of records in the dataset were found by the automated software scripts having problems with the five most crucial data quality parameters, meaning that the process of assessing and improving data before offering e-services is mandatory for all government institutions.

## 6.5 The e-Kosovo Model for Integrating e-Services in Electronic Platforms

The main objective or the goal is to apply the suitable model for integrating created e-Services in appropriate electronic platforms, such as the government portal of Kosovo. This process can be done after evaluating and enhancing the quality of data in the datasets.

Architecture of software can be defined as the basic framework of a program that defines its functionality and technical specifications. This is the main reason why picking the right architecture during the first stages of software development is crucial if we want our platform to be robust and functional for a long time [77].

According to Google's assessment of current software trends, the most often used method is with Microservice Architecture [78].

Using the most up-to-date methodology and tools for designing complex and sensitive systems, the three-tier architecture of the E-Kosovo electronic site was created to provide electronic services to citizens through a single point. The most crucial components of the whole design are the data's quality and the system's performance. Large corporations all throughout the world adopt the same approaches, including Microsoft and Google.

The "Domain Driven Design" (DDD) methodology and the "Model View Controller" web application development technology are the ones that were utilized to create this platform. This architecture is an improved one of "Multi-Tier Architecture," which has been enlarged

to accommodate new developing methodologies, component separation, such as Microservices, and software testing. Microservices, which are more narrowly focused on the distributed systems, are replacing monolithic architectural systems in the most recent creations of extremely complex systems [79].

This architecture consists of the latest techniques/ properties like:

- Dependency Injection
- Repositories
- Interfaces
- Services.

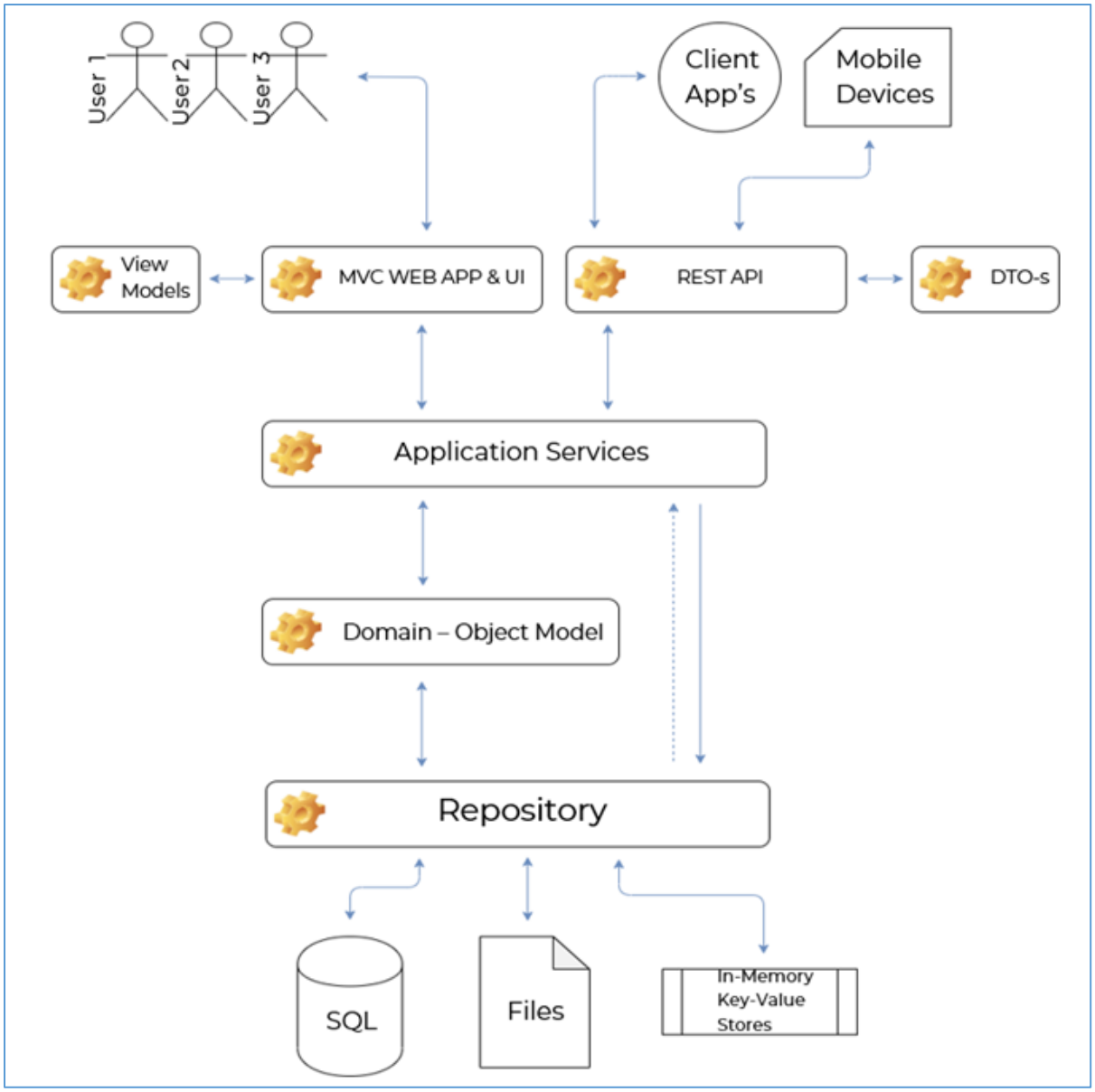



Figure 36: Domain Driven Design.

The following layers make up the technique/model [Figure 37] on which the development of web-based applications in.NET Core/.NET Framework is based:

- Model
- View
- Controller.

Each layer's tasks are detailed below:

- “Model”: All completed modeling and application logic are contained in this layer
- “View”: This layer displays how the program looks and how the data that the server or database receives are presented visually
- “Controller”: The interactions between the View and Model levels are controlled by this layer. Therefore, the logic of how application users interact with databases is regulated at this layer, as is the transfer of data from the server or database to the end user.

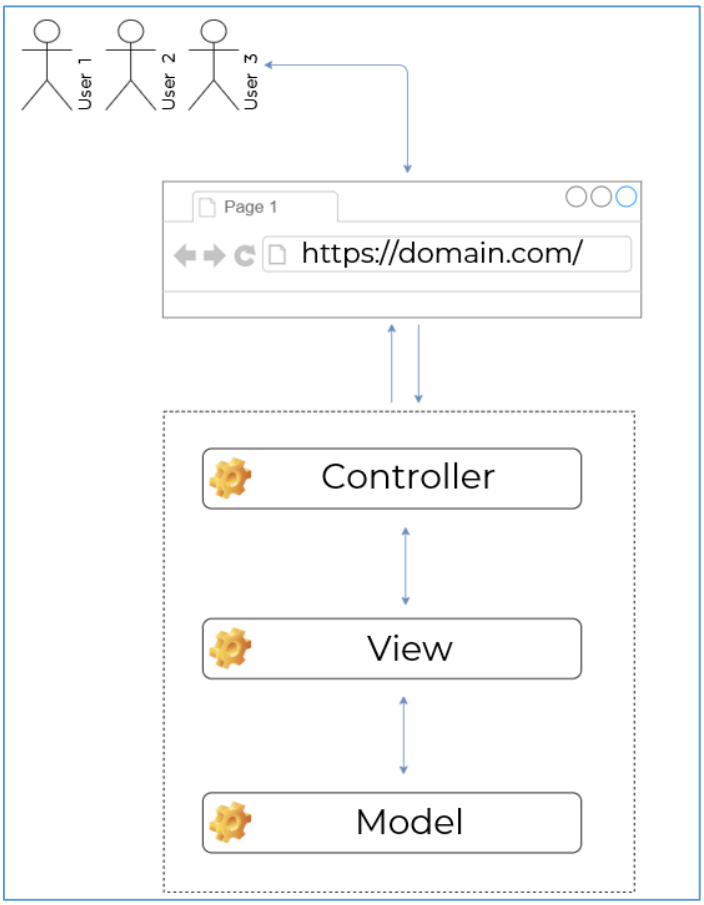



Figure 37: Controller View Model layers.

Figure 38 presents the physical organization of the integration of document issuance and verification from an electronic platform, such as the Kosovo government portal.

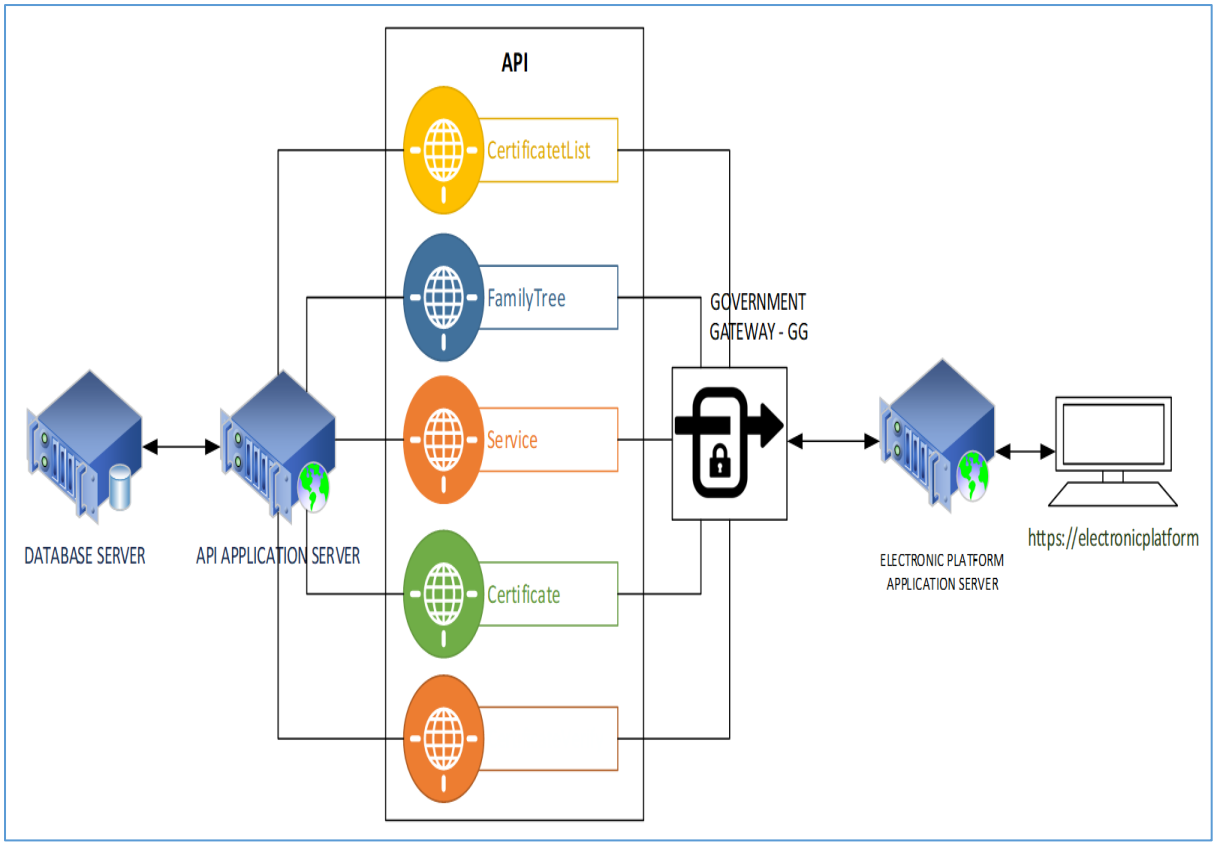



Figure 38: Physical structure for e-service integration.

Several models are generated and functionalized with the aim of integrating e-Services through government portals. We created the diagram that is shown below, which we believe is the most appropriate one to use when integrating services that deal with providing personal documents through official government platforms.

A model of a flowchart for integrating e-Services in the electronic platform is presented in Figure 39 below.

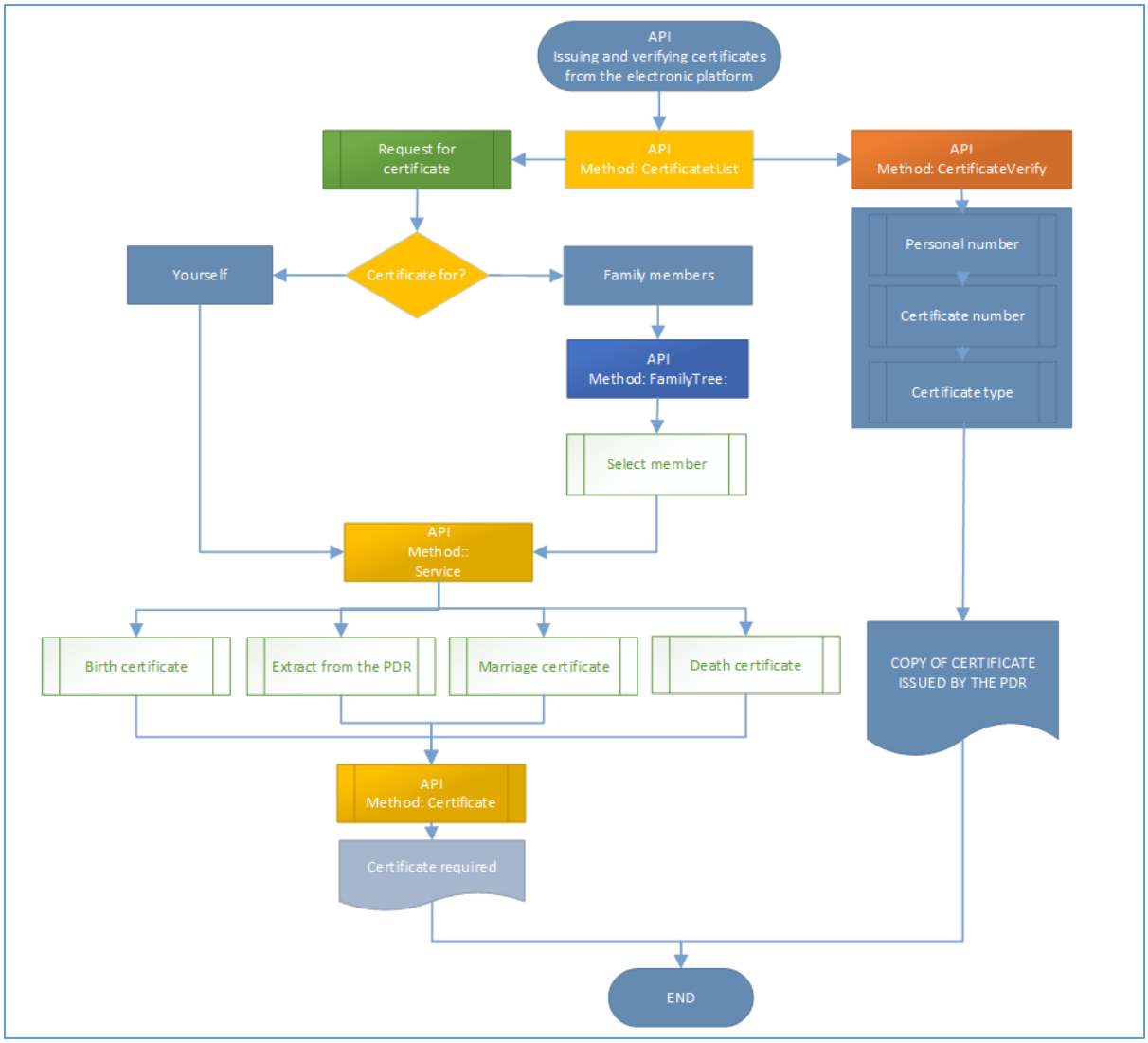



Figure 39: e-Services integration flowchart in the electronic platform.

A Signed Token is delivered from the Server to the Client through the API in each conceivable request to support token-based authentication.

Token Based Authentication is implemented using the JWT standard.

The most recent methods/features used in this architecture include:

1. The client enters User and Password to request authentication at the platform's endpoint of the API
2. The API for Token Based Authentication verifies the credentials supplied from the client by comparing them to those stored in the database
3. If step #2 for authentication is successful, the API will provide the client the token. The necessary data for identifying the client and the token's period of validity will be

on the token. Prior to transmission, these details are signed, and they are then sent to the client through HTTPS

4. All subsequent requests are made using the token, which the client stores in their device after obtaining it from the server via an API. In this way, the risk of user and password exposure is minimized

5. Upon receiving the token, the server checks to make sure it is correct before returning the required resources; if not, the client must start at step 1.

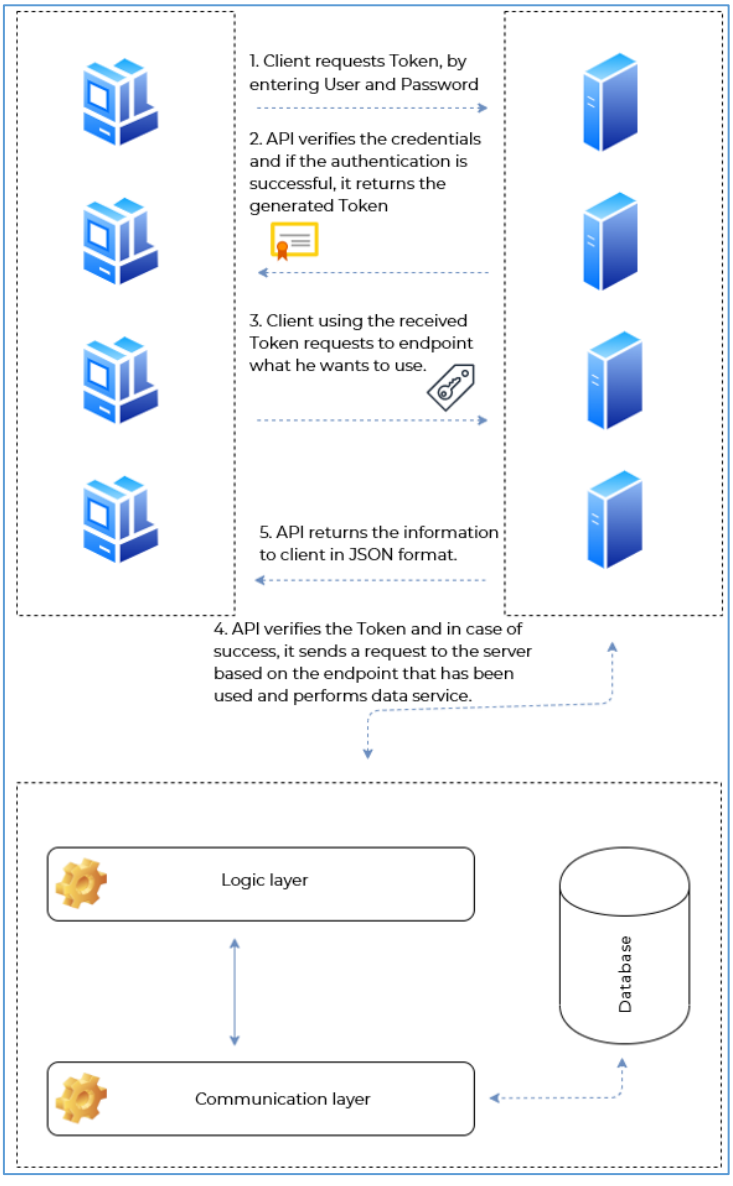



Figure 40: Authentication Using Tokens.

The API used by the government portal to issue and validate documents is described in the script below.

Access to API methods is made possible by clicking on the link where the input parameters and the desired outcome in jSon are, as shown below:

**API for document issuance and verification**

```
1:    Input parameters
2:      {
3:          "LanguageID": "sample string 1" – Language Parameters: depends
                          how many
4:      }
5:    Expected Result in jSon
6:    [
7:      {
8:          "DocumentID": "sample string 1",
9:          "Description": "sample string 2",
10:         "IssuedBy": "sample string 3", Depends from which platform the
                        documents are issued True=Yes and False=No
11:         "Verified": "sample string 4" Can be verified through electronic
                        platform
12:     },
13:     {
14:         "DocumentID": "sample string 1",
15:         "Description": "sample string 2",
16:         "IssuedBy": "sample string 3",
17:         "Verified": "sample string 4"
18:     }
19:   ]
```

## 6.6 Summary

In this chapter, through examination of five data quality aspects, we demonstrated data quality assessment and improvement of datasets. Measures including correctness, completeness, timeliness, consistency, and uniqueness are used to evaluate quality of the

data.

We have also created a special model for integrating e-Services in electronic platforms like for instance the Kosovo government portal after evaluating the dataset's quality and adjusting them. We first offered a flowchart showing how to do the process of integrating e-Services into the electronic platform and then we displayed an implementation solution using jSon. Finally, we discussed the physical structure and the process of integration of the issuance and verification of services.

Through this chapter, we also emphasized the significance of evaluating and enhancing data quality as a necessary step before putting into practice a specific model for offering e-Services through the Kosovo government portal.

Additionally, all of the findings from this study can help stakeholders provide professional and satisfying e-Services by minimizing issues and challenges from a technical and operational approach.

# 7
# Conclusion and Future Work

This chapter summarizes our research contributions for assessment of the impact of data quality for improvement of e-services in government institutions as well as presents future work.

## 7.1 Conclusion

Delivering improved e-services for citizens that can be offered through a government portal have been made possible by the data quality evaluation and enhancement in key datasets of governmental entities. The most effective method for assessing these advantages is to measure how satisfied individuals are with the e-services offered to them.

Through this dissertation, we emphasize the significance of data quality evaluation and improvement as a necessary step before the implementation of a particular model for providing e-Services through the government portal of Kosovo.

We also presented data quality frameworks overviewed by surveying and comparing types and structure of the data. Row, component, and information product data types are the data types that are treated. Structured, semi-structured, and unstructured data are the data structures that are examined. Additionally, we analyzed and compared the data quality dimensions utilized for particular frameworks, where some dimensions are found in several frameworks while others are found in just one. During assessment and improvement process of the data, we saw that accuracy, completeness, and timeliness are the most crucial dimensions. Different strategies and methodologies are introduced in this dissertation where each of these strategies employ specific steps to describe, analyze and enhance quality of the data. This process emphasizes the significance of evaluating and

enhancing data quality by selecting relevant frameworks and dimensions in order to raise the quality of datasets and services.

In this dissertation, by using data quality dimensions, we also provided an evaluation of the personal records data quality in assessed datasets. Datasets containing personal data are treated through several and appropriate dimensions such as completeness, accuracy and consistency, etc.

After assessing the datasets regarding data quality, we applied a variety of data matching algorithms to connect personal records from different data sources. The conclusion is that the Levenshtein Distance algorithm is the best algorithm for the matching and linking process when the focus in on quality of the data and in performance. Findings in this dissertation emphasize the significance of assessing DQ, and highlights the importance of identifying algorithm that suites the most for linking and matching process.

We first converted treated datasets to a suitable format and then began analyzing these data in order to execute the cleansing process of data in datasets and compare them from various and credible resources. After utilizing Order Dependencies Violation to identify abnormal data, we looked into the weekend and holiday Delay Reports problem and found the anomalous is specific data point or in all timeframe. In this dissertation, the most significant types of mistakes discovered on three official COVID-19 datasets were assessed which are still the main source for researchers all around the world. Therefore, even little mistakes in the official COVID-19 datasets might have a big impact on the numerous scientific outcomes. This situation demonstrates how crucial it is for COVID-19 official datasets to be accurate, updated, and completed in order to provide better scientific models and interpretations. Through the use of Power BI software, we presented the cumulative mortalities and cumulative cases retrieved from data collected in the datasets of WHO, JHU, and ECDC for Kosovo and North Macedonia, allowing users to notice the discrepancies in the data.

Through examination of five data quality dimensions, in this dissertation we treated data quality issues through data quality assessment and enhancement of the personal records dataset and car registration records. Datasets were evaluated on criteria like correctness, completeness, timeliness, consistency, and uniqueness. After evaluating and improving the

dataset's quality, we built a specific model for integrating e-Services into online platforms like the Kosovo government portal. We first offered a flowchart for the integration of e-Services in the electronic platform, and then we defined the physical structure of the integration of the issuance and verification services. Finally, we demonstrated how the proposed solution was implemented in practice.

## 7.2 Future Work

To further improve the solutions proposed in this dissertation, we have identified several areas for the future work. Based on experimental results, these results can guide stakeholders to offer professional and satisfactory e-Services through eliminating problems and obstacles from technical and operational perspective related to quality of the data.

For the assessment and improvement of data quality approach, what can be further investigated is the development of an accurate model that selects the best data quality dimensions that are necessary to be implemented with the aim that the datasets will be considered qualitative and ready to generate e-services for citizens by using them through a government portal. In this dissertation, we used dimensions such as: accuracy, completeness, consistency, uniqueness and timeliness and the challenge is to use only dimensions that are necessary, concrete and fit for the sets of data that will be assessed and improved.

Another direction that can be investigated in terms of assessing and improving data quality is improving models for interconnecting data from multiple resources that do not have unique fields through which they could be connected in order to provide more qualitative electronic services. In this sense, applying the proper algorithms for comparing the approaches for matching and connecting massive datasets of personal records from various sources has a substantial influence on enhancing the quality of data.

In the context of using appropriate algorithm for matching and linking data from multiple resources, many algorithms are used and also many improvements are done in these algorithms with the aim to have better and adequate results. There are still challenges in using and improving algorithms such as the Levenshtein distance algorithm and Damerau Levenshtein distance algorithm with fewer steps and less time to have qualitative data as

output. The search space between two datasets is attempted to be smaller by selecting the appropriate blocking variable. Creating pairings for comparison only among those who have the potential to be matches, such as those within same are or those who were born on specified dates, etc., tries to demonstrate that the effort and purpose are to narrow the search area.

The process of selecting the most suitable framework for evaluating and enhancing data quality presents additional challenges since various datasets require different frameworks depending on their structure, such as whether they are structured, semi-structured, or unstructured. Finding the best framework is a significant problem since it significantly decreases the amount of work necessary to produce datasets of sufficient quality for offering relevant e-services to citizens.

**REFERENCES**


[1] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," In: Data Sci. J., Volume 14, Issue 2, pp. 1–10, 2015.

[2] L. Li, "Data Quality and Data Cleaning in Database Applications," In: PhD thesis, School of Computing, September 2012.

[3] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso and P. Angeletti, "Improving Data Quality in Practice: A Case Study in the Italian Public Administration," In: Distrib. Parallel Databases, Volume 13, Issue 2, pp. 135-160, 2003.

[4] N. Askham, D. Cook, M. Doyle, H. Fereday, M. Gibson, U. Landbeck, R. Lee, C. Maynard, G. Palmer and J. Schwarzenbach, "The six primary dimensions for data quality assessment," In: Proc. DAMA U.K. Workshop Group, 2013.

[5] J.Vaughan, "Data Quality," In: https://www.techtarget.com/searchdatamanagement/definition/data-quality, Accessed July 2022.

[6] K. Peffers, T. Tuunanen, M. Rothenberger, M. and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," In: Journal of Management Information Systems, Volume 24, Issue 3, pp. 45-77, 2007.

[7] A. Ali, "Getting Started with Data Quality Services of SQL Server 2012 Using SSIS," In: https://www.mssqltips.com/sqlservertip/2593/getting-started-with-data-quality-services-of-sql-server-2012-using-ssis-part-4/, Accessed July 2022.

[8] M. Scannapieco, T. Catarci, "Data quality under a computer science perspective," In: Archivi & Computer, Volume 2, pp. 1-15, 2002.

[9] Y.W. Lee, D.M. Strong, B.K. Kahn and R.Y. Wang, "AIMQ: A Methodology for Information Quality Assessment," In: Information and Management, Volume 40, pp. 133-46, 2002.

[10] Centers for Disease Control and Prevention, "The Six Dimensions of EHDI Data Quality Assessment," In: https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf, Accessed July 2022.

[11] "Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model," In: ISOIEC 250122008, International Organisation for Standardization ISO, 2008.

[12] J. Tan, "How to improve data quality for machine learning?" In: https://towardsdatascience.com/how-to-improve-data-preparation-for-machine-

learning-dde107b60091, Accessed July 2022.

[13] L. Xingsen, Z. Lingling, Z. Peng and S. Yong, “Problems and Systematic Solutions in Data Quality,” In: International Journal of Services Sciences, Volume 2, Issue 1, pp. 53–69, 2009.

[14] M. Aljumaili, “Data Quality Assessment: Applied in Maintenance,” In: PhD thesis, Lulea University of Technology, 2016.

[15] H. Chen, D. Hailey, N. Wang, and P. Yu, “A review of data quality assessment methods for public health information systems,” In: Int. J. Environ. Res.Public Health, Volume 11, Issue 5, pp. 5170–5207, 2014.

[16] O. Azeroual and M. Abuosba, “Improving the Data Quality in the Research Information Systems,” In: (IJCSIS) International Journal of Computer Science and Information Security, Volume 15, Issue 11, 2017.

[17] E. Rahm and H.H. Do, “Data cleaning: problems and current approaches,” In: IEEE Data Engineering Bulletin, Volume 23, Issue 4, pp. 3–13, 2000.

[18] F. Naumann and U. Leser, “Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen,” In: Dpunkt Verlag, 1. Edition, October 2007.

[19] T. O’Brien, M. Helfert and A. Sukumar, “Classifying costs and effects of poor Data Quality – examples and discussion,” In: Annual Conference of Irish Academy of Management, Ireland, 2012.

[20] P. Woodall, M. Oberhofer and A. Borek, “A classification of data quality assessment and improvement methods,” In: Int. J. Inf. Quality, Volume 3, Issue 4, pp. 298-321, 2014.

[21] P. Woodall, A. Borek and A. Parlikad, “Data Quality Assessment: The Hybrid Approach,” In: Information & Management, Volume 50, Issue 7, pp.369–382, 2013.

[22] R. Y. Wang, “A product perspective on total data quality management,” In: Commun. ACM, Volume 41, Issue 2, pp. 58–65, 1998.

[23] N. Babar, “The Levenshtein Distance Algorithm” In: https://dzone.com/articles/the-levenshtein-algorithm-1, Accessed July 2022.

[24] R. Haldar and D. Mukhopadhyay, “Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach” In: Computing Research Repository-CORR, 2011.

[25] OpenGenus IQ, “Damerau Levenshtein distance” In: https://iq.opengenus.org/damerau-levenshtein-distance, Accessed July 2022.

[26] R. Wagner, and M. Fisher, “The string to string correction problem,” In: JACM,

Volume 21, Issue 1, pp. 168–173, 1974.

[27] C. Zhao and S. Sahni, “String correction using the damerau-levenshtein distance,” In: BMC Bioinf., Volume 20, Issue 11, 2019.

[28] M. Rehman, V. Esichaikul, and M. Kammal, “Factors influencing e-government adoption in Pakistan,” in: Transforming Government: People, Process and Policy, Volume 6, Issue 3, pp. 258-282, 2012.

[29] T. H. AlBalushi, “E-Services quality: A perspective of service providers and service users,” In: Digital Service Platforms E-Service IntechOpen, 2021.

[30] H. Anjoga, S. Nyeko and M. Kituyi, “A framework for usability of e-Government services in developing countries,” In: Journal of Accounting and Auditing: Research and Practice, pp. 1-15, 2017.

[31] C. Mkude, M. Wimmer, “E-government Systems Design and Implementation in Developed and Developing Countries: Results from a Qualitative Analysis,” In: 14th International Conference on Electronic Government (EGOV), Greece, pp.44-58, 2015.

[32] D. Kettani, M. Gurstein and A. El Mahdi, “Good governance and e-government: applying a formal outcome analysis methodology in a developing world context,” In: International Journal of Electronic Governance, Volum 2, Issue 1, pp. 22 – 54, 2009.

[33] M. A. Salam, “E-Governance for Good Governance through Public Service Delivery: an Assessment of District E-Service Centers in Bangladesh,” In: MA in Governance and Development, Institute of Governance Studies, 2013.

[34] S.F.H. Zaidi and M.K. Qteishat, “Assessing e-Government Service Delivery (Government to Citizen),” In: International Journal of eBussiness and eGovernment Studies, Volume 4, Issue 1, pp. 45-54, 2012.

[35] G. Mahlangu and E. Ruhode, “Towards a multidimensional model for assessing e-government service gaps,” In: South African Journal of Information Management, Volume 22, Issue 1, pp. 1–8, 2020.

[36] John Hopkins University and Medicine, Coronavirus Resource Center, In: https://coronavirus.jhu.edu/, July 2022.

[37] B. Xu, M. U. Kraemer, B. Gutierrez, S. Mekaru, K. Sewalk, A. Loskill et al., “Open access epidemiological data from the COVID-19 outbreak,” In: The Lancet Infectious Diseases, 2020.

[38] World Health Organization, “Framework and standards for country health information systems,” Second edition, 2012.

[39] V. Vasudevan, A. Gnanasekaran, V. Sankar, S. A. Vasudevan and J. Zou, “Disparity

in the quality of COVID-19 data reporting across India," In: BMC Public Health, Volume 21, Issue 1, pp. 1-12, 2021.

[40] World Health Organization, "Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update," In: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/, Accessed July 2022.

[41] European Centre for Disease Prevention and Control, "Data on the geographic distribution of COVID-19 cases worldwide," In: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide, Accessed July 2022.

[42] The National Institute of Public Health of Kosovo, https://www.facebook.com/IKSHPK/, Accessed July 2022.

[43] The Government of the Republic of North Macedonia, https://koronavirus.gov.mk/stat, Accessed July 2022.

[44] G. Wang, Z. Gu, X. Li, S. Yu, M. Kim, Y. Wang, L. Gao, and L. Wang, "Comparing and integrating US COVID-19 daily data from multiple sources: A county-level dataset with local characteristics," arXiv preprint arXiv:2006.01333, 2020.

[45] N. Altieri, R. L. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y. S. Tan, T. Tang, Y. Wang, C. Zhang, and B. Yu, "Curating a COVID-19 data repository and forecasting countylevel death counts in the united states," arXiv preprint arXiv:2005.07882, 2020.

[46] A. Ashofteh and J. M. Bravo, "A study on the quality of novel coronavirus (COVID-19) official datasets," Stat. J. IAOS, Volume 36, Issue 2, pp. 291-301, 2020.

[47] D. M. Strong, Y. W. Lee and R. Y. Wang, "Data Quality in Context," In: Communications of the ACM, Volume 40, Issue 5, pp. 103-110, 1997.

[48] P. Christen, "Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection," In: Springer Data Centric Systems and Applications, pp. 163-184, 2012.

[49] A. V. Levitin and T.C. Redman, "Data as a Resource: Properties, Implications, and Prescriptions," In: MIT Sloan Management Review, Volume 40, Issue 1, pp. 89-101, 1998.

[50] R.L. Leitheiser, "Data quality in health care data warehouse environments," In: HICSS, pp. 1-2, 2001.

[51] P. H. S. Panahy, F. Sidi, L. S. Affendey, and M. A. Jabar, "The impact of data quality dimensions on business process improvement," In: WICT, pp. 70–73, 2014.

[52] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data

consumers," In: Journal of Management Information Systems, Volume 12, Issue 4, pp. 5-33, 1996.

[53] Global Software Consultancyogic, "The Levenshtein Algorithm," In: https://www.cuelogic.com/blog/the-levenshtein-algorithm, Accessed July 2022.

[54] J. M. Jensen II, "Damerau-Levenshtein Edit Distance Explained," In: https://www.lemoda.net/text-fuzzy/damerau-levenshtein/, Accessed July 2022.

[55] A. Haug, F. Zachariassen and D. V. Liempd, "The costs of poor data quality," In: Journal of Industrial Engineering and Management, Volume 4, Issue 2, pp. 168-193, 2011.

[56] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A survey on data quality: Classifying poor data," In: Proc. IEEE Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 179–188, 2015.

[57] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," In: ACM Comput. Surv., Volume 41, Issue 3, 2009.

[58] F. Sidi, A. Ramli, M.A. Jabar, L.S. Affendey, A. Mustapha and H. Ibrahim, "Data quality comparative model for data warehouse," In: International Conference on Information Retrieval & Knowledge Management, pp. 268-272, 2012.

[59] A.F. Karr, A.P. Sanil, D.L. Bank, "Data quality: A statistical perspective," In: Statistical Methodology, Volume 3, Issue 2, pp. 137-173, 2006.

[60] D. McGilvray, "Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information," In: Morgan Kaufmann Publishers Inc., 2008.

[61] S. Abiteboul, P. Buneman, and D. Suciu, "Data on the Web: From Relations to Semistructured Data and XML," In: Morgan Kaufmann Publishers Inc., 2000.

[62] P. Buneman, "Semistructured data," In: Proceedings of the 16th ACM Symposium on Principles of Database Systems (PODS), pp. 117-121, 1997.

[63] D. Calvanese, D. Giacomo, and M. Lenzerini, "Modeling and querying semi-structured data," In: Networking and Information Systems Journal, Volume 2, Issue 2, pp. 253-273, 1999.

[64] C. Cichy and S. Rass, "An overview of data quality frameworks," In: IEEE Access, Volume 7, pp. 24634–24648, 2019.

[65] A. Giannoccaro, G.G. Shanks, and P. Darke, "Stakeholder perceptions of data quality in a data warehouse environment," In: Australian Computer Journal, Volume 31, Issue 4, pp. 110-116, 1999.

[66] E. D. Quality, "The data quality benchmark report," In: Experian Data Quality, Tech., 2015.

[67] T.C. Redman, “The impact of poor data quality on the typical enterprise,” In: Communications of the ACM, Volume 41, Issue 2, pp. 79–82, 1998.

[68] Y. Cui, J. Widom, and J.L. Wiener, “Tracing the lineage of view data in a warehousing environment,” In: Database Group, Stanford University, 1997.

[69] B. Heinrich, M. Kaiser, and M. Klier, “Metrics for measuring data quality foundations for an economic oriented management of data quality,” In: International Conference on Software and Data Technologies (ICSOFT), INSTICC/Polytechnic Institute of Setubal Barcelona, 2007.

[70] South Africa, “E-Government Portal,” In: https://www.eservices.gov.za/tonkana/home.jsf, Accessed July 2022.

[71] J. Sánchez-Torres, and I. Miles, “The role of future-oriented technology analysis in e government: A systematic review,” In: European Journal of Futures Research, Volume 5, pp. 1–18, 2017.

[72] S. Dawes, “Governance in the digital age: A research and action framework for an uncertain future,” In: Government Information Quarterly, Volume 26, pp. 257-264, 2009.

[73] S.K. Aikins, and D. Krane, “Are Public officials obstacles to citizen-centered e-government? An examination of municipal administrators' motivations and actions,” In: State and Local Government Review, Volume 42, pp. 87-103, 2010.

[74] E. Larsen, And R. Lee, “The Rise of the e-citizen: How People Use Government Agencies' Web Sites,” In: https://www.pewresearch.org/internet/2002/04/03/the-rise-of-the-e-citizen-how-people-use-government-agencies-web-sites, Accessed July 2022.

[75] W. I. M. Alabdallat, “Toward mandatory public e-services in Jordan,” In: Cogent Business and Management, Volume 7, Issue 1, pp. 120-131, 2020.

[76] A. Al-Taher, “E-government: Between theory and practice,” Amman: Araya Hall, 2009.

[77] K. Gos and W. Zabierowski, “The Comparison of Microservice and Monolithic Architecture,” In: IEEE XVIth International Conference on the Perspective Technologies and Methods in MEMS Design, pp. 150-153, 2010.

[78] Google Trends, “Microservices,” In: https://trends.google.com/trends/explore?date=now%201-d&q=microservices, Accessed July 2022.

[79] J. Stubbs, W. Moreira and R. Dooley, “Distributed systems of microservices using docker and serfnode,” In: Proceedings of the IEEE 2015 7th International Workshop on Science Gateways, pp. 34–39, 2017.

[80] H.N. Abdulkhudhur, I.Q. Habeeb, Y.Yusof and Sh.A.M Yusof, “Implementation of improved Levenshtein algorithm for spelling correction word candidate list generation”, In: Journal of Theoretical and Applied Information Technology, Volume 88. Issue 3, 2016.

[81] Z. P. Zhao, Z. M. Yin, Q. P. Wang, X. Z. Xu and H. F. Jiang, "An improved algorithm of Levenshtein Distance and its application in data processing", In: Journal of Computer Applications, Volume 29, pp. 424-426, 2009.

[82] S. Rani and J. Singh, "Enhancing Levenshtein's Edit Distance Algorithm for Evaluating Document Similarity", In: International Conference on Computing Analytics and Networks, pp. 72-80, 2017.

[83] H. Müller and J.C. Freytag, “Problems, Methods, and Challenges in Comprehensive Data Cleansing,” In: Technical Report HUB-IB-164, Humboldt University, Berlin, 2003.

[84] E. Rahm, and H.H. Do, “Data Cleaning: Problems and Current Approaches,” In: IEEE Data Engineering Bull., Volume 23, Issue 4, pp. 3-13, 2000.

## APPENDIX A

Source code for implementation of Levenshtein Algorithm:

```
FUNCTION [dbo].[LevenshteinDistance](@s1 nvarchar(100), @s2 nvarchar(100))
RETURNS int
AS
BEGIN
        if (len(@s2) = 0)
                begin
                        return @s2
                end
  DECLARE @s1_len int, @s2_len int, @i int, @j int, @s1_char nchar, @c int, @c_temp
int,
    @cv0 varbinary(8000), @cv1 varbinary(8000)
  SELECT @s1_len = LEN(@s1), @s2_len = LEN(@s2), @cv1 = 0x0000, @j = 1, @i = 1, @c = 0
  WHILE @j <= @s2_len
    SELECT @cv1 = @cv1 + CAST(@j AS binary(2)), @j = @j + 1
  WHILE @i <= @s1_len
  BEGIN
    SELECT @s1_char = SUBSTRING(@s1, @i, 1), @c = @i, @cv0 = CAST(@i AS binary(2)), @j
= 1
    WHILE @j <= @s2_len
    BEGIN
      SET @c = @c + 1
      SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j-1, 2) AS int) +
        CASE WHEN @s1_char = SUBSTRING(@s2, @j, 1) THEN 0 ELSE 1 END
      IF @c > @c_temp SET @c = @c_temp
      SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j+1, 2) AS int)+1
      IF @c > @c_temp SET @c = @c_temp
      SELECT @cv0 = @cv0 + CAST(@c AS binary(2)), @j = @j + 1
    END
    SELECT @cv1 = @cv0, @i = @i + 1
  END
  RETURN @c
END
```

---

```
Declare @Rowcount INT = 1;

WHILE (@Rowcount > 0)
BEGIN
INSERT INTO [dbo].[dmTable1Matching] WITH (TABLOCK)
     (
      [Register_ID]
```

```
        ,[CitizenID]
        ,[Name]
        ,[Surname]
        ,[DataOfBirth]
        ,[BirthPlace]
        ,[FathersName]
        ,[FathersSurname]
        ,[MothersName]
        ,[MothersSurname]
        ,[nBlockVariableID]
        ,[dtDataMatching])

SELECT
top(1000000)
l.Regjister_id,
c.CitizenID,
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.Emri)),LTRIM(RTRIM(c.vcName))),100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.Mbiemri)),LTRIM(RTRIM(c.vcSurname)))
,100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](dbo.udf_GetNumeric(convert(nvarchar(10),l.DataLindje
Convert,103)),dbo.udf_GetNumeric(convert(nvarchar(10),c.dtBirthdate,103))),100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.VendLindjaAdresa)),LTRIM(RTRIM(c.vcB
irthPlace))),100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.BEmri)),LTRIM(RTRIM(c.vcNameOfFath
er))),100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.BMbiemri)),LTRIM(RTRIM(c.vcSurname
OfFather))),100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.NEmri)),LTRIM(RTRIM(c.vcNameOfMot
her))),100),
ISNULL([dbo].[GetPercentageOfTwoStringMatchingLD](LTRIM(RTRIM(l.NMbiemri)),LTRIM(RTRIM(c.vcSurname
OfMother))),100),
1,
GETDATE()
from tblTable2 l , dbo.Table3 c
WHERE
(l.DataLindjeConvert)=(c.dtBirthdate)
and  not exists (select 1 from dmTableMatching dm where l.Regjister_id = dm.Register_ID and c.CitizenID =
dm.CitizenID )
OPTION (MAXDOP 16)

update top(1000000) [dbo].[dmTableMatching]
set [TotalPercentage] = Round(([Name] + [Surname] + [DataOfBirth] + [BirthPlace] + [FathersName] +
[FathersSurname] + [MothersName] +[MothersSurname] )/ 8,2)
where isnull([TotalPercentage],0) = 0
OPTION (MAXDOP 16)

SET @Rowcount = @@ROWCOUNT;

CHECKPOINT;   --<-- to commit the changes with each batch
END
```

---

```
USE [database1]
GO
```

```
/****** Object:  UserDefinedFunction [dbo].[GetPercentageOfTwoStringMatchingLD]    Script Date: 7/7/2022
1:56:49 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

-- select  [dbo].[GetPercentageOfTwoStringMatchingLDIndex]('17/11/1989','07/07/2022', 120)
alter FUNCTION [dbo].[GetPercentageOfTwoStringMatchingLDIndex]
(
    @string1 NVARCHAR(100) -- fusha nga dataset 1
  ,@string2 NVARCHAR(100) -- fusha nga dataset 2
  ,@index int           -- pesha e fushes
)
RETURNS float
AS
BEGIN

   DECLARE @levenShteinNumber float
   DECLARE @percentageOfGoodCharacters float
   DECLARE @string1Length float = LEN(@string1)
   , @string2Length float = LEN(@string2)
   DECLARE @maxLengthNumber float = CASE WHEN @string1Length > @string2Length THEN @string1Length
ELSE @string2Length END

               IF (@string1 = '' OR @string2 = '')
               BEGIN
               set @percentageOfGoodCharacters  = 0
               END
               ELSE
               BEGIN
   SELECT @levenShteinNumber =[dbo].[LevenshteinDistance](   @string1  ,@string2)

   DECLARE @percentageOfBadCharacters float = @levenShteinNumber * @index /
NULLIF(@maxLengthNumber,0)
   set @percentageOfGoodCharacters  = Round(@index - @percentageOfBadCharacters,2)
               END
               IF (@string1 = '' and @string2 = '')
               BEGIN
               set @percentageOfGoodCharacters = 0
               end
   -- Return the result of the function
   RETURN @percentageOfGoodCharacters
END
```

## APPENDIX B

Source code for implementation of Levenshtein Algorithm (improved in weight of fields):

```
Declare @Rowcount INT = 1;
WHILE (@Rowcount > 0)
BEGIN
INSERT INTO [dbo].[ dmTable1Matching] WITH (TABLOCK)
        (
         [Register_ID]
        ,[CitizenID]
        ,[Name]
        ,[Surname]
        ,[DataOfBirth]
        ,[BirthPlace]
        ,[FathersName]
        ,[FathersSurname]
        ,[MothersName]
        ,[MothersSurname]
        ,[nBlockVariableID]
        ,[dtDataMatching])
SELECT
top(1000000)
l.Regjister_id,
c.CitizenID,
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.Emri)),LTRIM(RTRIM(c.vcName)),1
20),120), -- Emri nga dataset 1, me emrin nga dataset 2, pesha 120
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.Mbiemri)),LTRIM(RTRIM(c.vcSurna
me)),120),120), -- Mbiemri nga dataset 1, me Mbiemri nga dataset 2, pesha 120
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](dbo.udf_GetNumeric(convert(nvarchar(10),l.DataL
indjeConvert,103)),dbo.udf_GetNumeric(convert(nvarchar(10),c.dtBirthdate,103)),120),120),-- DataLindje nga
dataset 1, me DataLindje nga dataset 2, pesha 120
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.VendLindjaAdresa)),LTRIM(RTRIM(
c.vcBirthPlace)),120),120),          -- VendLindja nga dataset 1, me VendLindja nga dataset 2, pesha 120
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.BEmri)),LTRIM(RTRIM(c.vcNameOf
Father)),80),80),                    -- BabaiEmri nga dataset 1, me BabaiEmri nga dataset 2, pesha 80
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.BMbiemri)),LTRIM(RTRIM(c.vcSurn
ameOfFather)),80),80),               -- BabaiMbiemri nga dataset 1, me BabaiMbiemri nga dataset 2, pesha 80
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.NEmri)),LTRIM(RTRIM(c.vcNameOf
Mother)),80),80),                    -- NenaEmri nga dataset 1, me NenaEmri  nga dataset 2, pesha 80
ISNULL([dbo].GetPercentageOfTwoStringMatchingLDIndex](LTRIM(RTRIM(l.NMbiemri)),LTRIM(RTRIM(c.vcSurn
ameOfMother)),80),80),               -- NenaMbiemri nga dataset 1, me NenaMbiemri nga dataset 2, pesha 80
1,
GETDATE()
from tblTable2 l , dbo.Table3 c
WHERE
```

```
 (l.DataLindjeConvert)=(c.dtBirthdate)
and  not exists (select 1 from dmDataMatching dm where l.Regjister_id = dm.Register_ID and c.CitizenID =
dm.CitizenID )
OPTION (MAXDOP 16)

update top(1000000) [dbo].[ dmTable1Matching]
set [TotalPercentage] = Round(([Name] + [Surname] + [DataOfBirth] + [BirthPlace] + [FathersName] +
[FathersSurname] + [MothersName] +[MothersSurname] )/ 8,2)
where isnull([TotalPercentage],0) = 0
OPTION (MAXDOP 16)

SET @Rowcount = @@ROWCOUNT;
CHECKPOINT;   --<-- to commit the changes with each batch
END
```

---

```
USE [DBImproved]
GO
/****** Object:  UserDefinedFunction [dbo].[GetPercentageOfTwoStringMatchingLD]    Script Date: 7/7/2022
1:56:49 PM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
-- select  [dbo].[GetPercentageOfTwoStringMatchingLDIndex]('17/11/1989','07/07/2022', 120)
alter FUNCTION [dbo].[GetPercentageOfTwoStringMatchingLDIndex]
(
    @string1 NVARCHAR(100) -- fusha nga dataset 1
  ,@string2 NVARCHAR(100) -- fusha nga dataset 2
  ,@index int           -- pesha e fushes
)
RETURNS float
AS
BEGIN

   DECLARE @levenShteinNumber float
   DECLARE @percentageOfGoodCharacters float
   DECLARE @string1Length float = LEN(@string1) , @string2Length float = LEN(@string2)
   DECLARE @maxLengthNumber float = CASE WHEN @string1Length > @string2Length THEN @string1Length
ELSE @string2Length END

              IF (@string1 = '' OR @string2 = '')
              BEGIN
              set @percentageOfGoodCharacters  = 0
              END
              ELSE
              BEGIN
   SELECT @levenShteinNumber =[dbo].[LevenshteinDistance](   @string1  ,@string2)

   DECLARE @percentageOfBadCharacters float = @levenShteinNumber * @index /
NULLIF(@maxLengthNumber,0)
   set @percentageOfGoodCharacters  = Round(@index - @percentageOfBadCharacters,2)
              END
              IF (@string1 = '' and @string2 = '')
              BEGIN
```

```
                set @percentageOfGoodCharacters = 0
        end
  -- Return the result of the function
  RETURN @percentageOfGoodCharacters
END
```

## APPENDIX C

Source code for implementation of Levenshtein Algorithm (improved in similar letter of Albanian alphabet):

```
USE [DBIMPROVED]
GO
/****** Object:  UserDefinedFunction [dbo].[LevenshteinDistanceN]    Script Date:
8/15/2022 9:44:49 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
--  SELECT [dbo].[LevenshteinDistanceN] ('qerim', 'çerim' )
CREATE FUNCTION [dbo].[LevenshteinDistanceN](@s1 nvarchar(100), @s2 nvarchar(100))
RETURNS float
AS
BEGIN
        if (len(@s2) = 0)
                begin
                        return @s2
                end
  DECLARE @s1_len int, @s2_len int, @i int, @j int, @s1_char nchar, @c int, @c_temp
float, @n bit, @nk int, @k int,
    @cv0 varbinary(8000), @cv1 varbinary(8000)
  SELECT @s1_len = LEN(@s1), @s2_len = LEN(@s2), @cv1 = 0x0000, @j = 1, @i = 1, @c =
0, @nk = 0, @k = 1
  WHILE @j <= @s2_len
    SELECT @cv1 = @cv1 + CAST(@j AS binary(2)), @j = @j + 1

  WHILE @i <= @s1_len

  BEGIN

    SELECT @s1_char = SUBSTRING(@s1, @i, 1), @c = @i, @cv0 = CAST(@i AS binary(2)), @j
= 1

    WHILE @j <= @s2_len
    BEGIN

      SET @c = @c + 1

      SET @c_temp = (CAST(SUBSTRING(@cv1, @j+@j-1, 2) AS int) +
        CASE WHEN @s1_char = SUBSTRING(@s2, @j, 1) THEN 0 ELSE 1 END)
      IF @c > @c_temp SET @c = @c_temp
      SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j+1, 2) AS int)+1
      IF @c > @c_temp SET @c = @c_temp

      SELECT @cv0 = @cv0 + CAST(@c AS binary(2)), @j = @j + 1

    END
```

```
    SELECT @cv1 = @cv0, @i = @i + 1


            WHILE @k <= @s1_len

        BEGIN
            set @n = 0;
            set @n = [dbo].[neighbors](SUBSTRING(@s1, @k, 1),SUBSTRING(@s2, @k, 1))
            if @n = 1 begin set @nk = @nk + 1 end
            set @k = @k + 1

            end
  END
  RETURN  @c - (@nk * 0.4)
END



USE [DBIMPROVED]
GO
/****** Object:  UserDefinedFunction [dbo].[LevenshteinDistance]    Script Date:
8/15/2022 9:44:42 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
--  SELECT [dbo].[neighbors] ('a', 'q' )
CREATE FUNCTION [dbo].[neighbors](@s1 nvarchar(100), @s2 nvarchar(100))
RETURNS bit
AS
BEGIN

declare @c bit
set @c=case when @s1 in ('e') and @s2 in ('ë') then 1 else
       case when @s1 in ('ë') and @s2 in ('e') then 1 else
            case when @s1 in ('i') and @s2 in ('j','y') then 1 else
            case when @s1 in ('j') and @s2 in ('i','y') then 1 else
            case when @s1 in ('y') and @s2 in ('i','j') then 1 else
            case when @s1 in ('q') and @s2 in ('ç') then 1 else
            case when @s1 in ('ç') and @s2 in ('q') then 1 else
            0
                  end end end end end end end

  RETURN @c
END
```

## APPENDIX D

Source code for implementation of Levenshtein Algorithm (improved in edit distance value, for Replacement (substitution)=0.6, Insert=1.2, Delete=1.2):

```
USE [DBIMPROVED]
GO
/****** Object:  UserDefinedFunction [dbo].[LevenshteinDistance]    Script Date:
8/17/2022 10:08:26 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
--  SELECT [dbo].[LevenshteinDistanceN2] ('betimi', 'bekim' )
alter FUNCTION [dbo].[LevenshteinDistanceN2](@s1 nvarchar(100), @s2 nvarchar(100))
RETURNS float
AS
BEGIN
        if (len(@s2) = 0)
                begin
                        return @s2
                end
  DECLARE @s1_len int, @s2_len int, @i int, @j int, @s1_char nchar, @c int, @c_temp
float, @n bit, @nk int, @k int,
    @cv0 varbinary(8000), @cv1 varbinary(8000)
  SELECT @s1_len = LEN(@s1), @s2_len = LEN(@s2), @cv1 = 0x0000, @j = 1, @i = 1, @c =
0, @nk = 0, @k = 1
  WHILE @j <= @s2_len
    SELECT @cv1 = @cv1 + CAST(@j AS binary(2)), @j = @j + 1

  WHILE @i <= @s1_len

  BEGIN

    SELECT @s1_char = SUBSTRING(@s1, @i, 1), @c = @i, @cv0 = CAST(@i AS binary(2)), @j
= 1

    WHILE @j <= @s2_len
    BEGIN

      SET @c = @c + 1

      SET @c_temp = (CAST(SUBSTRING(@cv1, @j+@j-1, 2) AS int) +
        CASE WHEN @s1_char = SUBSTRING(@s2, @j, 1) THEN 0 ELSE 1 END)
      IF @c > @c_temp SET @c = @c_temp
      SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j+1, 2) AS int)+1
      IF @c > @c_temp SET @c = @c_temp

      SELECT @cv0 = @cv0 + CAST(@c AS binary(2)), @j = @j + 1

    END


    SELECT @cv1 = @cv0, @i = @i + 1
```

```
        declare @ins int =0, @del int = 0

        if @s1_len > @s2_len
        begin
        set @del = @s1_len - @s2_len
        end
        else
        begin
        set @ins =  @s2_len - @s1_len
        end


  END
  RETURN  (@c * 0.6) +  (@ins * 1.2) +  (@del *1.2) - (@ins * 0.6) - (@del * 0.6)
END
```